Monotone missing data and repeated controls of fallible authors

Raats, V.M.

Publication date:
2004

Document Version
Publisher's PDF, also known as Version of record

Monotone Missing Data and Repeated Controls of Fallible Auditors

Monotone Missing Data and Repeated Controls of Fallible Auditors

PROEFSCHRIFT

ter verkrijging van de graad van doctor aan de Universiteit van Tilburg, op gezag van de rector magnificus, prof.dr. F.A. van der Duyn Schouten, in het openbaar te verdedigen ten overstaan van een door het college voor promoties aangewezen commissie in de aula van de Universiteit op

vrijdag 10 december 2004 om 10.15 uur

door

VERA MARIA RAATS

geboren op 24 juni 1976 te Zundert.

PROMOTOR: prof. dr. B.B. van der Genugten
COPROMOTOR: dr. J.J.A. Moors

To my mother and grandmother

Acknowledgements

Now my time as a Ph.D. student has come to an end, it is good to have the opportunity to thank the people who contributed (directly or indirectly) to this thesis. First of all, my supervisors Ben van der Genugten and Hans Moors for their kindness, enthusiasm and seemingly endless patience. Thanks also to the other members of the committee: Paul van Batenburg, John Einmahl, Jan Magnus, Ton Steerneman and Marleen Willekens, for the time and effort spent on reading the thesis.

On the non-scientific side I have greatly appreciated the support of my friends and family. I would like to mention a few of them separately. Ten eerste tante Nel en ome Sjef voor een thuis en eigenlijk teveel dingen om op te noemen, maar vooral omdat ze de heerlijke, liefdevolle mensen zijn die ze zijn. Secondly, Will Princen for being an excellent mirror. Last, but certainly not least, Steffan, for his love and support, and for making me laugh!

VERA RAATS
SEPTEMBER 2004, LONDON

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation

By the time this thesis was started in 2000, six companies were responsible for the social security payments in the Netherlands. Together, they paid more than € 22 billion a year on sickness and unemployment benefits, and the like. Although they were, to a large extent, independent and self-regulating, they were under twofold inspection: they were subject to external auditors[1]' assessments of their annual financial statements, and a supervising institution - called the CTSV (nowadays the IWI) - produced annual assessments of the legality of their payments on behalf of the Department of Social Security. Furthermore, internal audit departments performed extensive tests on randomly selected payments, the results of which were shared with both external auditors and CTSV.

These checks were useful, since Dutch social security rules and regulations were (and are) notoriously complicated. Mistakes and misinterpretations therefore were easily made, even by experts in the field. According to the annual report 2003 of IWI , the incorrect payments in that year - although only 1.6% of the total sum paid - amounted to a huge € 365 million. Table 1.1.1 - taken from the annual report 2002 of IWI (in Dutch) - contains some detailed information about social security payments in earlier years. The first column of the table mentions

[1]Throughout this thesis we use the term "audit" (and similarly "auditor") in its general meaning of inspections (executed for example by controllers, surveyors or accountants)'.

different kinds of social security payments; for example, the Wajong was meant for disabled adolescents and students.

	Payments 2002 (in million €)	Percentage errors 2002	2001
WAO	12011	0.2	0.2
WAZ	584	4.5	1.2
Rea	693	5.4	1.9
ZA	1124	9.1	2.0
BIA	8	7.0	2.0
Wajong	1584	0.9	0.7
Wazo	856	3.8	4.2
TW	287	6.3	2.1
WW	3939	4.6	2.9

Table 1.1.1: Social security payments

One of the methods that the CTSV used to check for incorrect payments and incorrect assessments of the internal auditors, is double checking. So, after the auditors had checked the book values of a large number of sampled records, this supervising organization double checked a subsample of these records to assess the quality of the auditors' work. For some records the CTSV's judgement would differ from the auditors'. Although this did not necessary imply an auditor's error since the difference maybe caused by different interpretation of the payment rules, we will use the term "error" throughout this thesis. Since the CTSV had great expertise, it assumed that their own check is faultless. So we ended up with a sample of single checked records (with only the fallible assessment) plus a sample of double checked records from which we can compare the number and size of the errors found by the auditor with the true errors discovered by the expert. The question remained how to combine the information from both the fallible auditor and expert to draw the most accurate conclusions about the true errors in the population.

This thesis tries to answer this question by the statistical modeling and inference of repeated audit controls. In a formal repeated audit control a fallible auditor checks a random sample of records. A subsample of these (already checked)

records is checked again by another (more skillful) auditor. This procedure may be repeated several times until the final auditor, considered to be infallible, gives the true values of some sampled records which have already been checked by all previous auditors.

Repeated audit controls are related to missing data problems. Standard statistical methods usually analyse a number of variables, observed for a fixed number of cases. However, it frequently occurs that not all of the data entries are observed for all cases, implying that some data entries are missing; these missing data problems occur frequently in practice and have received a lot of attention in the literature. Repeated audit controls can be regarded as missing data problems. For example, in case of two rounds, the expert's judgement is observed for the double checked records, but it is missing for the single checked records for which only the (fallible) auditor's assessment is available.

Though we formulate the problem in terms of a fallible and an infallible auditor, it is important to note that our analysis is also valid for the general quality control problem in which objects are classified by a (cheap) error-prone device and a random subsample is classified again by a precise (but expensive) device to adjust for misclassification. Finally, it is also important to note that the problem of fallible auditors is not only relevant for the Dutch social security payments. The last couple of years this has been shown only too often by (extreme) cases like Enron and Worldcom which made it into the global news.

1.2 Outline

In this thesis several models for repeated audit controls will be discussed. They differ with respect to the number of fallible auditors and the kind of variables (categorical, continuous or a mixture). Chapter 2 starts with the case from which our research originated: the repeated control of the Dutch social security payments (involving only one fallible auditor plus the expert). Since the parameter of interest is the fraction of incorrect payments, the auditor and expert classify a record as either correct or incorrect, leading to dichotomous variables. The corresponding classification probabilities are important additional parameters.

The model of Chapter 2 was first introduced by Tenenbein (1970) and has recently also been studied by Barnett *et al.* (2001). Both papers mainly focussed on point estimation (and in particular maximum likelihood estimation). Since in auditing practice upper limits usually are at least as important as point estimates, we discuss two approaches to determine upper limits for the fraction of incorrect records in the population: a numerical procedure to determine classical upper confidence limits (which is a generalization of Moors *et al.* (2000)) and the Bayesian approach. It is shown that the classical approach leads to very conservative upper limits; the Bayesian upper limits are in general lower.

Chapter 3 presents a general framework for repeated audit controls with categorical variables and/or several fallible auditors; the model of Chapter 2 is the simplest situation within this setting. We study two different sampling methods: stratified and random sampling. In stratified sampling, previous classification results determine the next sample sizes for all classifications separately, while in random sampling they only determine the total sample size for the next auditor. Stratified sampling is often applied in practice. We derive the maximum likelihood estimators for both methods and propose a solution for maximum likelihood estimators which are not uniquely defined, a frequently occurring problem in practice. We compare three different approaches to derive upper limits, including the Bayesian approach. Our Bayesian model deviates essentially from a previously adopted Bayesian model: the prior distributions are formulated for a different, more natural, set of parameters. The underlying independence assumptions of our approach seem to be more realistic than the usual ones. To determine the Bayesian upper limit, we make use of the data augmentation algorithm of Tanner and Wong (1987) for determining Bayesian posterior distributions in missing data problems.

So, in these two chapters models for repeated audit controls with categorical variables were analysed; in the remaining chapters models for continuous variables, and a mixture of categorical and continuous variables will be treated. These models are highly relevant in practice, since often one is not only interested in the fraction of errors in the population, but also in the total size of the errors.

Chapters 4 and 5 discuss multivariate linear regression with monotone missing observations of the - continuous - dependent variables; the latter means that

the dependent variables can be ordered in such a way that if an observation of a dependent variable for a record is missing, the observations of all subsequent dependent variables for the same record are also missing. See Schafer (1997) *e.g.* for a more extensive discussion about monotone data patterns. The explanatory variables are assumed to have been completely observed: for these variables no missing observations occur. This model is an important generalization of the case with just the constant as explanatory variable, which has received a lot of attention in the literature (see Bhargava (1962) *e.g.*). Note that the multivariate regression model with monotone missing observations is widely applicable, repeated audit controls being only one example. In case of a repeated audit control, the dependent variables are the (fallible) auditors' and the expert's judgement; the known book value (and the constant) act as the explanatory variables.

In Chapter 4 we derive closed form expressions for the least squares and maximum likelihood estimators using projections, these estimators get a clear geometrical interpretation. The existing iterative method for calculating maximum likelihood estimates in missing data problems, is the widely used EM-algorithm, which numerically converges to the maximum likelihood estimates. In comparison, our method has two advantages: the easy interpretation and the direct calculation which of course is much faster and more precise. We include (sets of) MANOVA-tables enabling us to perform exact likelihood ratio tests on the coefficients. They lead to a new type of distribution, a generalization of the well-known Wilks' distribution. Similar to the approximations for the Wilks' distribution for complete data (see Bartlett (1947) *e.g.*), several approximations for this generalized Wilks' distribution are derived and compared by simulation.

In Chapter 5 we look at several additional features of the multivariate regression model. First of all, we prove that the estimators of the previous chapter - and a more general class of estimators - are consistent. This result is used to prove the consistency of the iterative weighted least squares algorithm. For the sake of completeness the EM-algorithm for our model is given; it is similar but not identical to the one of Meng and Rubin (1993). A generalization of the model with just the constant as explanatory variable is obtained as a special case: one-way MANOVA.

It would not be realistic to assume a continuous model for the errors of the records since, in auditing practice, the errors often equal zero. However, if the errors are not zero they can take on a lot of different values. In the final Chapter 6 we use the models of the previous chapters to construct a more realistic model for repeated audit controls with a mixture of discrete and continuous variables. This model consists of a discrete submodel for the classification probabilities and a continuous submodel for the non-zero errors using conditional regression. We present the maximum likelihood estimators for the model parameters, and a new estimator for the mean size of the errors in the population. Simulation shows that this last estimator outperforms the estimators proposed by Barnett *et al.* (2001).

1.3 Publication background

The chapters in this thesis are chronologically ordered. They are based on previous publications which (almost all) have been written in cooperation with B.B. van der Genugten and J.J.A. Moors. Chapters 2, 3 and 4 can be read independently; Chapter 4 is necessary for understanding Chapter 5, while Chapter 6 demands knowledge of Chapter 2, 3 and 4.

The contents of Chapter 2 are derived from my Master's thesis which was written during an internship at Deloitte and Touche. The thesis was converted into research report Raats and Moors (2000) and published as Raats and Moors (2003). Chapter 2 coincides with Raats and Moors (2003) as published, except for the shortened introduction and some minor layout changes.

Chapter 3 has been published as Raats *et al.* (2004b) (with some minor layout changes) and consists of research report Raats *et al.* (2002a) and, additionally, the Bayesian approach for determining upper limits.

Chapter 4 is based on research reports Raats *et al.* (2002b) and Raats (2004). It is essentially a revised version of Raats *et al.* (2002b) with two additional sections: Section 4.4 about relative efficiency and Section 4.10 about the approximations of the generalized Wilks' distribution (which is a curtailed version of Raats (2004)).

Chapter 5 consists of Raats *et al.* (2004a) and two additional sections: Section 5.4 about the EM-algorithm and Section 5.5 about one-way MANOVA.

Chapter 6 is based on Raats *et al*. (2004). To avoid needless repetitions, the two underlying research reports of the last two chapters have been shortened considerably.

Chapter 2

Dichotomous data, two rounds

2.1 Introduction

As mentioned in Section 1.1, six companies are responsible for the social security payments in the Netherlands. For one of these six companies, an internal auditor reported 16 errors in a random sample of 500 payments, leading to an estimated error rate of 3.2% and a 95% upper confidence limit of 4.8%. The supervising CTSV decided to double check this result. Of the 500 payments evaluated by the auditor, a random subsample of 53 was checked once more - independently and error free - by an external auditor of the CTSV. The subsample contained two errors found by the auditor; both appeared to be true errors indeed. However, among the remaining 51 payments, approved by the internal auditor, the CTSV auditor found one additional error. The question now is how to derive from the information in both sample and subsample, point and interval estimates for the population error rate.

The problem recently received attention from two sides; besides, we found that it was discussed much earlier. A brief review of the relevant papers follows, going back in history; to present a detailed overview of recent developments, not only published papers, but also research reports are mentioned. The most recent published contribution is Barnett *et al*. (2001), based on the research report Barnett *et al*. (2000). It discusses the two type of mistakes an auditor may make:

- evaluating an incorrect payment as ‘correct’ (missing an error), and

- evaluating a correct payment as 'incorrect' (making up an error),

and presents the maximum likelihood estimator (MLE) for the population error rate. (Besides, a quantitative approach is followed: three methods are proposed to estimate the total population error from the *size* of the observed errors. The quantitative approach will be discussed in Chapter 6; for the moment, we will only be concerned with qualitative variables.)

The same MLE was derived in Moors (1999), and applied to the Dutch social security example in Raats and Moors (2000). The latter was based on the Master's thesis Raats (1999); it is a generalization of Moors *et al.* (2000) where only one type of auditor's mistake was considered: since no made up error was found in the CTSV subsample, the corresponding probability was put equal to 0 *a priori*. Further, a numerical method was given to find confidence intervals for the population error rate.

But neither Barnett *et al.* (2000) nor Moors (1999) can claim priority. Near the end of 2001 we discovered that the same MLE was already derived in Tenenbein (1970). Compare also Tenenbein (1971) and Tenenbein (1972). Besides, we found that this estimator can be easily derived as well from the more general monotone sampling approach, discussed by Little and Rubin (2002) (and (1987), the earlier edition).

This chapter is organized as follows. Sections 2.2 - 2.4 discuss the classical approach of repeated audit controls. Section 2.2 describes the repeated control model and sets out our notation. Section 2.3 briefly discusses the MLE's, in particular for the population error rate. In Section 2.4 a numerical method to determine a classical upper confidence limit for the error rate is presented; the method is illustrated by means of the CTSV example. However, we show that this classical confidence limit is very conservative, due to the presence of nuisance parameters; consequently, it is of limited practical use.

Therefore, it seems logical to follow the Bayesian approach. Section 2.5 presents a Bayesian model for the situation of just one possible auditor's mistake: (s)he may miss errors, but never makes them up. Section 2.6 contains the Bayesian approach for the extended model where both types of auditor's mistakes

may occur. The final Section 2.7 discusses the main results and gives some conclusions. Also, extensions in two different directions are briefly discussed.

2.2 The model

The model which we consider in this paper, coincides with the model which first was considered by Tenenbein (1970) and more recently by Barnett *et al*. (2001). However, we introduce another, more intuitive, notation that can easily be generalized for extended audit controls with categorical data and more than two rounds; see Chapter 3.

In the following notation the subindex 0 stands for incorrect and subindex 1 for correct. Consider a population in which a fraction p_0 of the records is incorrect. The (internal) auditor decides a randomly drawn record to be 'incorrect' or 'correct'. The quotation marks indicate a decision; the same phrases without them indicate the true situation. So we take the possibility that the auditor misclassifies the record into account: with (conditional) probability $p_{1|0}$ an incorrect record is (erroneously) judged to be 'correct' and with probability $p_{0|1}$ a correct record is misclassified as 'incorrect'.

From the three error probabilities

$$\begin{cases} p_0 & = Pr(\text{random record is incorrect}) \\ p_{1|0} & = Pr(\text{auditor misses an error}) \\ p_{0|1} & = Pr(\text{auditor makes up an error}) \end{cases} \tag{2.2.1}$$

other probabilities as the joint probability p_{10} (of a random record being correct and being misclassified as 'incorrect') can be derived. The number of records found to be 'correct' and 'incorrect' by the auditor in a random sample of size n_1 will be denoted by C_1 and C_0, respectively.

Now, an external auditor who is assumed to be faultless (the expert) checks a subsample of the records, of size n_2, once more. In this subsample the expert determines the true number C_{+0} of incorrect records; C_{00} of these errors were already found by the first auditor, but C_{10} were missed. Of the C_{+1} correct records in the subsample, C_{01} were misclassified as 'incorrect' by the first auditor, while the remaining C_{11} were correctly classified.

The $n_1 - n_2$ remaining records are checked only once; C_{0-} and C_{1-} denote the number of 'incorrect' and 'correct' values among them. Table 2.2.1 shows the complete information obtained from both checks.

	Total	Single checked sample	Double checked sample		
			Expert		
First auditor			Total	correct	incorrect
'correct'	C_1	C_{1-}	C_{1+}	C_{11}	C_{10}
'incorrect'	C_0	C_{0-}	C_{0+}	C_{01}	C_{00}
Total	n_1	$n_1 - n_2$	n_2	C_{+1}	C_{+0}

Table 2.2.1: Classification frequencies

It will appear to be helpful to introduce some more notation, in particular error probabilities, based on the auditor's judgements; compare the monotone missing data approach in Little and Rubin (2002). These inverse error probabilities are

$$\begin{cases} \pi_0 & = Pr(\text{'incorrect'}) \\ \pi_{1|0} & = Pr(\text{correct}|\text{ 'incorrect'}) \\ \pi_{0|1} & = Pr(\text{incorrect}|\text{ 'correct'}). \end{cases} \tag{2.2.2}$$

Figure 2.2.1 shows both sets of parameters in the double checked sample.

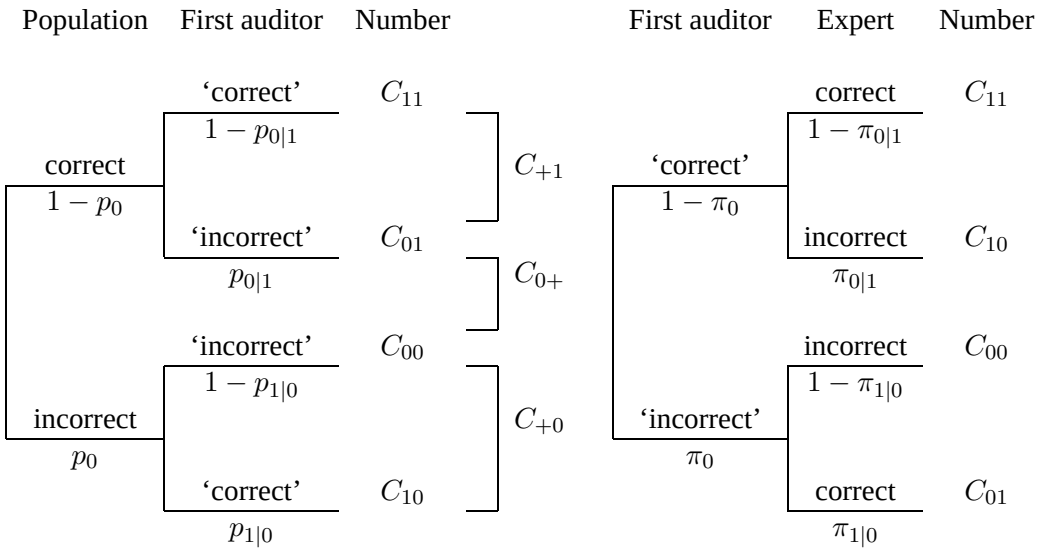


Figure 2.2.1: Classification frequencies and probabilities

Joint probabilities as π_{01} (a random record being classified as 'incorrect' by the auditor and as correct by the expert) $= p_{10}$ follow from these. Besides, the following one-to-one relations exist between (2.2.1) and (2.2.2):

$$\begin{cases} p_0 = (1-\pi_0)\pi_{0|1} + \pi_0(1-\pi_{1|0}), & \pi_0 = (1-p_0)p_{0|1} + p_0(1-p_{1|0}) \\ p_{1|0} = \dfrac{(1-\pi_0)\pi_{0|1}}{(1-\pi_0)\pi_{0|1} + \pi_0(1-\pi_{1|0})}, & \pi_{1|0} = \dfrac{(1-p_0)p_{0|1}}{(1-p_0)p_{0|1} + p_0(1-p_{1|0})} \\ p_{0|1} = \dfrac{\pi_0\pi_{1|0}}{(1-\pi_0)(1-\pi_{0|1}) + \pi_0\pi_{1|0}}, & \pi_{0|1} = \dfrac{p_0p_{1|0}}{(1-p_0)(1-p_{0|1}) + p_0p_{1|0}}. \end{cases} \tag{2.2.3}$$

Under the assumption of random sampling with replacement, all random variables in the model have (conditional) binomial distributions with the probabilities (2.2.2) as parameters:

$$\begin{cases} \mathcal{L}(C_0) &= B(n_1;\pi_0) \\ \mathcal{L}(C_{0+}|C_0 = c_0) &= B(n_2; c_0/n_1) \\ \mathcal{L}(C_{01}|C_{0+} = c_{0+}) &= B(c_{0+};\pi_{1|0}) \\ \mathcal{L}(C_{10}|C_{1+} = c_{1+}) &= B(c_{1+};\pi_{0|1}). \end{cases} \tag{2.2.4}$$

The likelihood is the product of these conditionally independent binomial distributions.

2.3 Estimation

From (2.2.4), MLE's for the parameter set (2.2.2) are found immediately; for the original set (2.2.1), they then follow directly from (2.2.3):

$$\begin{cases} \widehat{P}_0 &= \dfrac{C_1}{n_1}\dfrac{C_{10}}{C_{1+}} + \dfrac{C_0}{n_1}\dfrac{C_{00}}{C_{0+}} \\ \widehat{P}_{1|0} &= \dfrac{C_1}{n_1}\dfrac{C_{10}}{C_{1+}} \Big/ \left(\dfrac{C_1}{n_1}\dfrac{C_{10}}{C_{1+}} + \dfrac{C_0}{n_1}\dfrac{C_{00}}{C_{0+}}\right) \\ \widehat{P}_{0|1} &= \dfrac{C_0}{n_1}\dfrac{C_{01}}{C_{0+}} \Big/ \left(\dfrac{C_1}{n_1}\dfrac{C_{11}}{C_{1+}} + \dfrac{C_0}{n_1}\dfrac{C_{01}}{C_{0+}}\right). \end{cases} \tag{2.3.1}$$

The same expressions can be found in Tenenbein (1970), Moors (1999) and Barnett *et al.* (2001). The MLE's have clear interpretations, based on (2.2.3); furthermore, it is straightforward that the moment estimators coincide with the MLE's. Note that for $C_{01} = 0$, the formulae for $\widehat{P}_0$ and $\widehat{P}_{1|0}$ reduce to expression (6) in Moors (1999), treating the one error type situation with $p_{0|1} = 0$.

The estimator for our main parameter p_0 breaks down when either $C_{1+} = 0$ or $C_{0+} = 0$. Though this situation can be avoided by using stratified sampling such as Tenenbein (1970) remarked and the next chapter discusses in more detail, in case of random sampling these events can occur. In case of $C_{1+} = 0$ or $C_{0+} = 0$, the likelihood does not lead to a unique MLE and somewhat arbitrary values have to be chosen. Heuristic arguments (details can be found in Moors (1999)) lead to the following MLE for p_0 (compare also (3.5.2)):

$$\widehat{P}_0 = \begin{cases} \dfrac{C_{10}}{n_2} & \text{for } C_{0+} = 0 \\ \dfrac{C_1}{n_1}\dfrac{C_{10}}{C_{1+}} + \dfrac{C_0}{n_1}\dfrac{C_{00}}{C_{0+}} & \text{for } 0 < C_{1+} < n_2 \\ \dfrac{C_{00}}{n_2} & \text{for } C_{1+} = 0. \end{cases} \tag{2.3.2}$$

Appendix 2.8.1 shows that the distribution of (2.3.2) is symmetrical with respect to the point $(p_{1|0}, p_{0|1}) = (0.5, 0.5)$. The intuitive explanation is that for high values of the misclassification probabilities $p_{1|0}$ and $p_{0|1}$, all the auditor's judgements should be reversed: 'correct' is better interpreted as 'incorrect', and *vice versa*.

2.4 Upper limits

Following the argumentation of Cox and Hinkley (1974) Chapter 7, p. 229, it is straightforward that an $(1-\alpha)$ upper confidence limit for p_0, given a point estimate $\widehat{p}_0$, can be obtained from

$$p_0^u = \max_{p_0}\{p_0, p_{1|0}, p_{0|1} : Pr(\widehat{P}_0 \leq \widehat{p}_0 | p_0, p_{1|0}, p_{0|1}) \geq \alpha\}. \tag{2.4.1}$$

The calculation of the upper limit (2.4.1) is illustrated by means of the CTSV-example. Table 2.4.1 contains the numerical data of this practical example which was presented in Moors *et al.* (2000) and described in Section 2.1.

	Total	Single checked sample	Double checked sample		
			Expert		
First auditor			Total	correct	incorrect
'correct'	$c_1 = 484$	$c_{1-} = 433$	$c_{1+} = 51$	$c_{11} = 50$	$c_{10} = 1$
'incorrect'	$c_0 = 16$	$c_{0-} = 14$	$c_{0+} = 2$	$c_{01} = 0$	$c_{00} = 2$
Total	$n_1 = 500$	$n_1 - n_2 = 447$	$n_2 = 53$	$c_{+1} = 50$	$c_{+0} = 3$

Table 2.4.1: CTSV example

For this example, (2.3.1) results in the ML estimates

$$\widehat{p}_0 = 0.051, \quad \widehat{p}_{1|0} = 0.372, \quad \widehat{p}_{0|1} = 0.000.$$

To determine the accompanying 95% upper confidence limit p_0^u in (2.4.1), the quantity

$$p_0^u|p_{1|0}, p_{0|1} = \max_{p_0}\{p_0 : Pr(\widehat{P}_0 \leq 0.051|p_0, p_{1|0}, p_{0|1}) \geq 0.05\}$$

has to be calculated for all possible values of $p_{1|0}$ and $p_{0|1}$. Thanks to the symmetry of $\widehat{P}_0$ with respect to the point $(p_{1|0}, p_{0|1}) = (0.5, 0.5)$, the calculations may be limited to the $p_{0|1}$ interval [0, 0.5]. Figure 2.4.1 gives a 3-dimensional illustration.

Subsequently, the maximum of $p_0^u|p_{1|0}$,$p_{0|1}$ over all possible values of $p_{1|0}$ and $p_{0|1}$ has to be determined. This maximum was found to be 0.121; it was realized for $(p_{1|0}, p_{0|1}) = (0.914, 0.000)$ and - because of the symmetry - for $(p_{1|0}, p_{0|1}) = (0.086, 1.000)$. Note that the $p_{0|1}$ value 1 is inconsistent with the sample result $c_{11} = 50$ in Table 2.4.1; however, this is irrelevant since we are interested in the final $\widehat{p}_0$ value 0.051 and not in the individual classification numbers. The solid curve in Figure 2.4.2 shows $p_0^u|p_{1|0}$,$p_{0|1}$ for $p_{0|1} = 0$ and the accompanying maximum p_0^u; for comparison, this function is shown as well for $p_{0|1} = 0.3$ (the dotted curve).

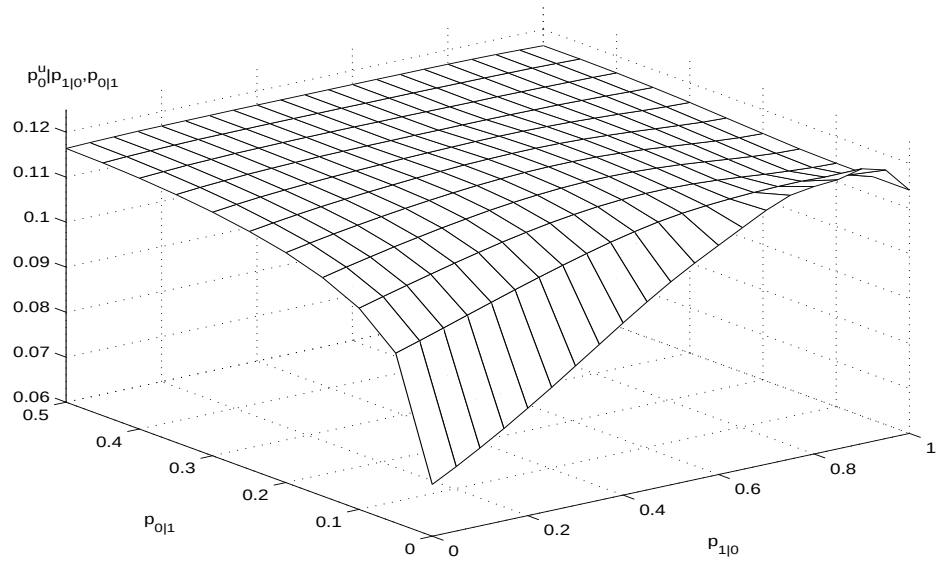


Figure 2.4.1: $p_0^u|p_{1|0}, p_{0|1}$ for $\widehat{p}_0 = 0.051$

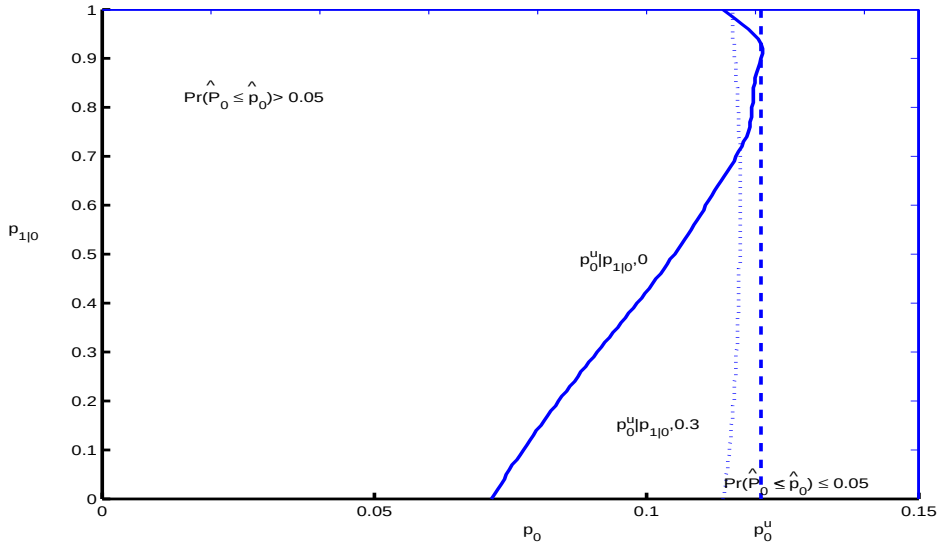


Figure 2.4.2: $p_0^u|p_{1|0}, p_{0|1}$ for $\widehat{p}_0 = 0.051; p_{0|1} = 0$ and $p_{0|1} = 0.3$

It is interesting to compare these results with the numerical findings in Moors *et al.* (2000). In the reduced model, the maximum likelihood (ML) estimates for p_0 and $p_{1|0}$ are still determined according to (2.3.1) and therefore coincide with the ML estimates of the extended model such as determined earlier. However, a slightly 95% lower upper confidence limit $p_0^u = 0.120$ was calculated.

In the present model - as in the reduced model - the upper limit is realized for a very high value of $p_{1|0}$ or $p_{0|1}$. In reality, such high values will not often occur and the upper limit (2.4.1) can be very conservative. This can also be concluded from Appendix 2.8.2, which contains the coverage of the 95% upper limits for different sets of parameters. The error probabilities and the first three sets of sample sizes coincide with the ones analysed by Barnett *et al.* (2001). In all these cases, the coverage of the classical upper limit (2.4.1) is at least 95%. The coverage is higher for the lower p_0-value. Furthermore, the results indicate that $p_{1|0}$ has a considerably larger impact on the coverage than $p_{0|1}$. The latter part of Appendix 2.8.2 is included to enable a comparison between the coverage of the Bayesian and classical upper limits in Section 2.7. In all cases, the coverage is calculated from simulation runs with 10,000 iterations each.

2.5 Bayesian approach for one error type

Different authors already discussed the Bayesian approach for fallible audits. Viana (1994) analysed a model with possible misclassifications but without a double check. York *et al.* (1995) presented the Bayesian approach for a double sampling scheme with two fallible auditors. Geng and Asano (1989) looked in more detail at the Bayesian model where some classifications of a fallible auditor are checked again by an infallible expert. However, they considered the situation with two dichotomous variables in each audit round, whereas our model only considers one dichotomous variable per round (the classification 'correct' or 'incorrect'). Moreover, Geng and Asano (1989) used Dirichlet priors for the inverse error probabilities (2.2.2) rather than for the (natural) model parameters (2.2.1). The latter was also done by Schafer (1997) who discussed the Bayesian approach for general multinomial, monotone missing data problems. In this and the next section

we will formulate the Bayesian model in terms of priors for the parameters (2.2.1). For simplicity, first the one error type model is considered.

In the one error type situation where $p_{0|1}$ (the probability of making up an error) is *a priori* set to zero as in Moors *et al.* (2000), the model contains two unknown parameters. In the Bayesian approach these two parameters p_0 and $p_{1|0}$ are viewed as realizations of random variables P_0 and $P_{1|0}$. Their prior distribution represents the researcher's knowledge before the sample results are obtained. A logical choice for the marginal prior distributions of P_0 and $P_{1|0}$ is the beta distribution, as the conjugated distribution of the binomial sample results. Further, independence of P_0 and $P_{1|0}$ (the quality of the population is independent of the quality of the auditor) seems reasonable, so that the joint prior distribution of P_0 and $P_{1|0}$ is the product of two beta distributions:

$$\mathcal{L}(P_0, P_{1|0}) \propto p_0^{\alpha_0-1}(1-p_0)^{\alpha_1-1} p_{1|0}^{\alpha_{1|0}-1}(1-p_{1|0})^{\alpha_{0|0}-1}.$$

The prior knowledge about p_0 ($p_{1|0}$) is reflected by the parameters α_0 and α_1 ($\alpha_{1|0}$ and $\alpha_{0|0}$).

In combination with the binomial sample results (2.2.4) this leads to the following joint posterior distribution of (P_0,$P_{1|0}$):

$$\begin{aligned}&\mathcal{L}(P_0, P_{1|0}|\text{sample results}) \propto \\ &\quad \textstyle\sum_{k=0}^{c_{1-}} \Big[(-1)^k \binom{c_{1-}}{k} p_0^{c_{+0}+c_{0-}+\alpha_0+k-1}(1-p_0)^{c_{+1}+\alpha_1-1}\cdot \\ &\qquad\qquad p_{1|0}^{c_{10}+\alpha_{1|0}-1}(1-p_{1|0})^{c_{00}+c_{0-}+\alpha_{0|0}+k-1}\Big].\end{aligned}$$

Integrating over $P_{1|0}$ gives the marginal posterior distribution of the main parameter P_0:

$$\begin{aligned}&\mathcal{L}(P_0|\text{sample results}) \propto \\ &\quad \textstyle\sum_{k=0}^{c_{1-}} \Big[(-1)^k \binom{c_{1-}}{k} p_0^{c_{+0}+c_{0-}+\alpha_0+k-1}(1-p_0)^{c_{+1}+\alpha_1-1}\cdot \\ &\qquad\qquad B(c_{10}+\alpha_{1|0}, c_{00}+c_{0-}+\alpha_{0|0}+k)\Big].\end{aligned} \tag{2.5.1}$$

with $B(a,b) = \int_0^1 x^{a-1}(1-x)^{b-1}dx$. Note that (2.5.1) is a weighted average of beta distributions with signed weights $(-1)^k \binom{c_{1-}}{k} B(c_{10}+\alpha_{1|0}, c_{00}+c_{0-}+\alpha_{0|0}+k)$.

As point estimate b_0 for p_0 in the Bayesian approach we take the mode of the marginal posterior distribution of P_0 since, in general, this corresponds to the ML

estimate when the prior distribution is uniform (see Little & Rubin (2002), p. 105 *e.g.*); the 0.95-quantile of the marginal posterior distribution is the Bayesian 95% upper limit b_0^u. Note that by integrating over $P_{1|0}$, all different values of $p_{1|0}$ are taken into consideration, and not only the worst values as in the classical approach. Hence, b_0^u will be lower than p_0^u in general.

An important feature of the Bayesian approach is the choice of the prior distribution parameters. In practice, prior information about p_0 could be obtained from previous audits of the same population. To get an idea of the quality of the fallible auditor, one could look at education, years of experience, performance in similar previous audits *et cetera*. However, since we do not have such information, the CTSV example will be analysed for the non-informative, or uniform, prior and some other hypothetical priors.

If no specific prior knowledge is available, all possible values of (P_0, $P_{1|0}$) can be considered as equally probable; this leads to the non-informative prior, defined by $\alpha_0 = \alpha_1 = \alpha_{1|0} = \alpha_{0|0} = 1$. The choice $\alpha_1 > \alpha_0$ *e.g.* reflects the researcher's belief that lower values of P_0 are more likely. For simplicity, $\alpha_0 = \alpha_{1|0} = 1$ will be chosen throughout; for α_1 and $\alpha_{0|0}$ the values 1 and 5 will be considered. The choice of this latter value is based on the following argumentation. If a record is randomly classified, the probability of a misclassification is 0.5. For a beta prior with parameters 1 and 5 the 95% upper limit is about 0.5. So the probability of misclassification is less than 0.5 with probability 0.95. Indeed, it seems not unreasonable to assume that classifications by a qualified auditor will outperform random classifications.

The Bayesian approach is now applied to the practical CTSV example. For the data in Table 2.4.1 and the non-informative prior, the posterior (2.5.1) becomes

$$\mathcal{L}(P_0|\text{sample results}) \propto \sum_{k=0}^{433} \left[(-1)^k \binom{433}{k} p_0^{17+k} (1-p_0)^{50} B(2, 17+k) \right].$$

Figure 2.5.1 shows this distribution; the Bayesian estimates b_0 and b_0^u are shown as well.

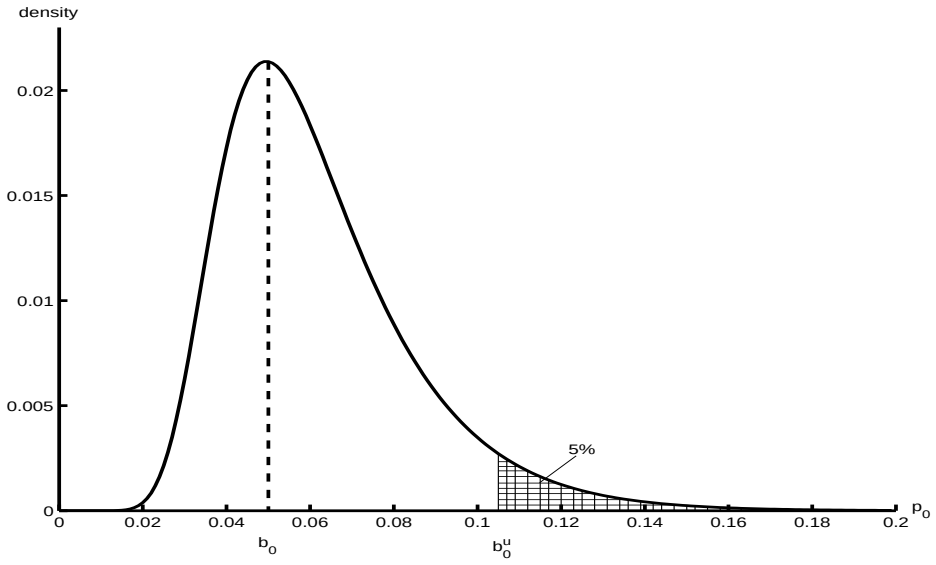


Figure 2.5.1: Marginal posterior distribution P_0; one error type

Table 2.5.1 summarizes these Bayesian estimates for four different priors; for comparison, the classical estimates, mentioned in Section 2.4, are added.

Parameters prior		Bayesian estimates	
α_1	$\alpha_{0\|0}$	b_0	b_0^u
1	1	.050	.105
5	1	.048	.101
1	5	.042	.075
5	5	.042	.073
Classical estimates		.051	.120

Table 2.5.1: Point estimates and upper limits for p_0 ; $\alpha_0 = \alpha_{1|0} = 1$

All Bayesian estimates are lower than the corresponding classical results. For the upper limits, this is caused by the additional information represented in the prior. Especially prior knowledge about the quality of the auditor has a large impact on the estimates; the researcher's belief that $p_{1|0}$ is low ($\alpha_{0|0} = 5$) leads to a considerable reduction of b_0 and b_0^u. The reason is that there is less sample

information concerning $p_{1|0}$ than p_0.

2.6 Bayesian approach for two error types

The model with two error types contains $p_{0|1}$ as a third unknown parameter. Independence of P_0 and ($P_{1|0}$, $P_{0|1}$) seems reasonable (the quality of the population is independent of the quality of the auditor), but independence of $P_{1|0}$ and $P_{0|1}$ is questionable. Nevertheless, this assumption is made here to simplify the calculations. Starting from marginal beta distributions, the joint prior distribution of P_0, $P_{1|0}$ and $P_{0|1}$ then reads:

$$\mathcal{L}(P_0, P_{1|0}, P_{0|1}) \propto p_0^{\alpha_0-1}(1-p_0)^{\alpha_1-1} p_{1|0}^{\alpha_{1|0}-1}(1-p_{1|0})^{\alpha_{0|0}-1} p_{0|1}^{\alpha_{0|1}-1}(1-p_{0|1})^{\alpha_{1|1}-1}. \tag{2.6.1}$$

In combination with the binomial sample results (2.2.4), this leads to the following joint posterior distribution:

$$\begin{aligned}&\mathcal{L}(P_0, P_{1|0}, P_{0|1}|\text{sample results}) \propto \\ &\quad p_{1|0}^{c_{10}+\alpha_{1|0}-1}(1-p_{0|1})^{c_{11}+\alpha_{1|1}-1} \sum_{j=0}^{c_{1-}} \sum_{k=0}^{c_{0-}+j} \Big[(-1)^j \tbinom{c_{1-}}{j} \tbinom{c_{0-}+j}{k} p_0^{c_{+0}+k+\alpha_0-1} \cdot \\ &\quad (1-p_0)^{c_{+1}+c_{0-}+j-k+\alpha_1-1}(1-p_{1|0})^{c_{00}+k+\alpha_{0|0}-1} p_{0|1}^{c_{01}+c_{0-}+j-k+\alpha_{0|1}-1}\Big].\end{aligned}$$

Integrating over the nuisance variables $P_{1|0}$ and $P_{0|1}$ leads to the marginal posterior distribution of the main parameter P_0 :

$$\begin{aligned}&\mathcal{L}(P_0|\text{sample results}) \propto \\ &\quad \sum_{j=0}^{c_{1-}} \sum_{k=0}^{c_{0-}+j} \Big[(-1)^j \tbinom{c_{1-}}{j} \tbinom{c_{0-}+j}{k} p_0^{c_{+0}+k+\alpha_0-1}(1-p_0)^{c_{+1}+c_0+j-k+\alpha_1-1} \cdot \\ &\quad B(c_{10}+\alpha_{1|0}, c_{00}+k+\alpha_{0|0}) B(c_{01}+c_{0-}+j-k+\alpha_{0|1}, c_{11}+\alpha_{1|1})\Big].\end{aligned} \tag{2.6.2}$$

Again, the marginal posterior distribution is the weighted average of beta distributions.

The Bayesian approach is applied to the example of Section 2.4. Using the non-informative prior in combination with the sample results in Table 2.4.1, (2.6.2)

can be simplified to:

$$\mathcal{L}(P_0|\text{sample results}) \propto \sum_{j=0}^{433} \sum_{k=0}^{14+j} (-1)^j \binom{433}{j} \binom{14+j}{k} p_0^{3+k} (1-p_0)^{64+j-k} B(2, 3+k) B(16+j-k, 50).$$

Figure 2.6.1 shows the marginal posterior distribution and the Bayesian estimates b_0 and b_0^u.

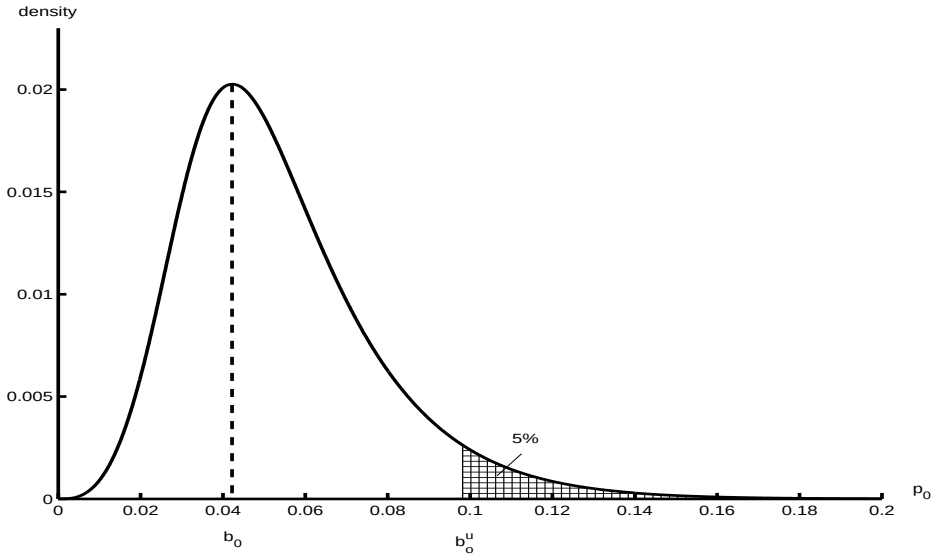


Figure 2.6.1: Marginal posterior distribution P_0; two error types

Table 2.6.1 contains the classical results calculated in Section 2.4 and the Bayesian results for eight different priors.

As in the situation with one error type, all Bayesian estimates are lower than the corresponding classical results and again prior knowledge about $p_{1|0}$ has a larger impact on the results than prior knowledge about p_0. Prior knowledge about $p_{0|1}$ hardly has any impact although this parameter, just like $p_{1|0}$, concerns the quality of the auditor. The explanation is that there is much more sample information on $p_{0|1}$: this parameter is estimated from the $c_{+1} = 50$ correct records in the double-checked sample, and $p_{1|0}$ from only the $c_{+0} = 3$ incorrect values.

Parameters prior			Bayesian estimates	
α_1	$\alpha_{0\|0}$	$\alpha_{1\|1}$	b_0	b_0^u
1	1	1	.042	.098
1	1	5	.042	.098
5	1	1	.041	.093
5	1	5	.041	.093
1	5	1	.036	.068
1	5	5	.036	.068
5	5	1	.035	.066
5	5	5	.035	.067
Classical estimates			.042	.0116

Table 2.6.1: Point estimates and upper limits for $p_0; \alpha_0 = \alpha_{1|0} = \alpha_{0|1} = 1$

As shown earlier the coverage of the classical $(1 - \alpha)$ upper limit often is (much) higher than $1 - \alpha$. Since the Bayesian upper limit is based more on the sample estimates of the nuisance parameters than the classical upper limit that considers the worst-case situation, the Bayesian coverage may be expected to be closer to $1 - \alpha$. Due to numerical difficulties caused by the signed weights, we only calculated Bayesian coverage for relatively small sample sizes. The last part of Appendix 2.8.2 shows our numerical results for non-informative priors. For these small sample sizes, there is not much difference between the coverage of the classical and the Bayesian upper limits.

2.7 Conclusions and further research

In this chapter both the classical approach and the Bayesian approach of two models for the repeated audit control have been discussed. The calculations were illustrated by means of the actual data from the Dutch CTSV-investigation. Table 2.7.1 shows some more results, for slightly different sample outcomes; the Bayesian results are based on the non-informative prior.

Model	n_1	n_2	c_0	c_{0-}	c_{+0}	c_{10}	c_{01}	*Classical* $\widehat{p}_0$	*Classical* p_0^u	*Bayesian* b_0	*Bayesian* b_0^u
Single check	500	-	16	-	-	-	-	.032	.048	.035	.048
Double check	500	53	-	14	2	0	-	.032	.092	.038	.077
one error type	500	53	-	14	3	1	-	.051	.120	.050	.105
Double check	500	53	-	14	3	1	0	.051	.121	.042	.098
two error types	500	53	-	14	3	1	1	.042	.116	.037	.094

Table 2.7.1: Classical and Bayesian point estimates and upper limits

The most striking feature of this table is that all double check models lead to increased upper limits; even if the expert finds not a single additional error (line 2) p_0^u and b_0^u are 90 and 60%, respectively, larger than when the auditor is assumed to be infallible (line 1).

Lines 3 an 4 represent the empirical data found in Dutch social security payments, where the first auditor made up no errors, but missed one error. In line 3 the model includes only the possibility of missing errors, in line 4 the possibility of making up errors is considered as well. Extending the model with this second error type has not much influence on the classical results, while the Bayesian estimates decrease. Of course, if the auditor made up one of the errors (line 5), all estimates decrease.

Appendix 2.8.3 contains some additional results for the different models. In this appendix the upper limits are only calculated for small sample sizes ($n_1 = 50, n_2 = 20$), since the calculations of the upper limits are rather time consuming and dramatically increase with sample sizes. The Bayesian 95% upper limits are calculated for the non-informative prior, as well as for the prior with one parameter set to 5 (and the other parameters set to 1).

Note that the Bayesian upper limits are generally smaller than the classical ones, although Table 2.8.4 shows two exceptions. This can be explained as follows, for example for the one error type situation. Introduce the Bayesian upper limit $b_0^u|p_{1|0}$ for a given value of $p_{1|0}$, analogously to $p_0|p_{1|0}$. Then $b_0^u|p_{1|0} < p_0^u|p_{1|0}$ will hold, unless the prior distribution of p_0 is concentrated around (much) higher values than the sample information. Now, b_0^u is obtained by averaging $b_0^u|p_{1|0}$ with respect to $p_{1|0}$, while $p_0^u = \max_{p_{1|0}}(p_0^u|p_{1|0})$ considers the worst case. Consequently,

only exceptionally b_0^u will exceed p_0^u; for the cases considered here, this will occur in particular for the non-informative prior.

Generalizations of the present model which are discussed in the next chapter, concern more audit rounds, categorical data, and stratified instead of random sampling.

The models discussed in this chapter consider rather elementary situations, that deviate from practical auditing conditions in two main respects.

- In practice, the total size of all errors will be of even greater importance than the error rate p_0: hence the size of individual errors will have to be taken into account. Barnett *et al*. (2001) presented a classical estimator for the mean size of the errors with a double sampling design. Chapter 4 presents estimation methods and algorithms for monotone missing continuous data which will be applied to repeated audit controls in Chapter 6. Laws and O'Hagan (2000) discussed the Bayesian model for a flawless sample check with taintings. A similar approach could be followed for the double sampling scheme.

- The previous research started from random sampling. However, in auditors' practices, selection with probabilities proportional to the recorded values ('monetary unit sampling' or MUS) is applied frequently. Hence, it would be interesting to investigate this sampling method as well.

In the Bayesian approach it was assumed that the probability of missing an error is independent of the probability of making up an error. Since this assumption is questionable, it would be interesting to repeat the above investigations without assuming independence. Following Gunel (1984), Dirichlet-beta priors could be used to incorporate dependence.

Finally, a number of more theoretical issues remain. For example, according to Lehmann and Casella (1998), p. 176, no uniformly most accurate confidence set will in general exist in the presence of nuisance parameters, as in our case, but perhaps our method of constructing upper limits can be improved.

2.8 Appendices

2.8.1 Symmetry of the MLE

In case of two possible error types, it will be shown here by means of three consecutive lemmas that the distribution of the MLE $\widehat{P}_0$ for p_0 is symmetric with respect to $(p_{1|0}, p_{0|1}) = (0.5, 0.5)$, that is: $\mathcal{L}(\widehat{P}_0|p_0, p_{1|0}, p_{0|1}) = \mathcal{L}(\widehat{P}_0|p_0, 1-p_{1|0}, 1-p_{0|1})$.

Introduce V =$(C_{+0}, C_{10}, C_{01}, C_{0-})$, define the functions $f : R^4 \to R^4$ and $h : [0,1]^3 \to [0,1]^3$ by

$$f(v) = f(c_{+0}, c_{10}, c_{01}, c_{0-}) = (c_{+0}, c_{+0} - c_{10}, n_2 - c_{+0} - c_{01}, n_1 - n_2 - c_{0-})$$

and

$$h(p) = h(p_0, p_{1|0}, p_{0|1}) = (p_0, 1 - p_{1|0}, 1 - p_{0|1}),$$

and define the set A_c for all $c \in [0,1]$ by

$$A_c = \{v : \widehat{p}_0(v) = c\}.$$

Note that $f = f^{-1}$ and $h = h^{-1}$.

Lemma 2.8.1. $f(A_c) = A_c$.

Proof. The special case $v = (c_{+0}, c_{+0}, 0, c_{0-})$ implies $f(v) = (c_{+0}, 0, n_2 - c_{+0}, n_1 - n_2 - c_{0-})$ and $\widehat{p}_0(v) = \widehat{p}_0(f(v)) = \dfrac{c_{+0}}{n_2}$. In the general case, $\widehat{p}_0(v) = \widehat{p}_0(f(v))$ can be proved similarly. Hence $v \in A_c$ implies $f(v) \in A_c$, and vice versa. □

Lemma 2.8.2. $Pr(V = v|p) = Pr(V = f(v)|h(p))$.

Proof. By direct verification, using (2.2.4). □

Lemma 2.8.3. $Pr(\widehat{P}_0 = c|p) = Pr(\widehat{P}_0 = c|h(p))$.

Proof.

$$\begin{array}{rcccl} Pr(\widehat{P}_0 = c|h(p)) = & Pr(V \in A_c|h(p)) & = & Pr(V \in f(A_c)|h(p)) \\ = & Pr(V \in A_c|p) & = & Pr(\widehat{P}_0 = c|p) \end{array}$$

where the second equality follows from Lemma 2.8.1 and the third from Lemma 2.8.2. □

2.8.2 Simulated coverage

Table 2.8.1 contains the simulated coverages of the 95% classical upper limits. In the last column, coverage of the Bayesian upper limit with a non-informative prior is given in parentheses.

Probabilities			$n_1 = 1000$,	$n_1 = 3000$,	$n_1 = 3000$,	$n_1 = 50$,
p_0	$p_{1\|0}$	$p_{0\|1}$	$n_2 = 100$	$n_2 = 100$	$n_2 = 300$	$n_2 = 20$
.10	.20	.011	99.8	99.9	99.7	100.0 (99.3)
.10	.20	.033	99.5	99.5	99.0	100.0 (99.6)
.10	.20	.056	99.2	99.2	98.3	100.0 (99.8)
.10	.60	.011	98.6	98.7	97.6	100.0 (98.6)
.10	.60	.033	98.2	98.3	96.6	100.0 (98.8)
.10	.60	.056	97.9	98.0	96.1	100.0 (99.4)
.20	.20	.025	99.6	99.6	99.6	97.1 (97.2)
.20	.20	.075	98.6	98.8	98.7	97.1 (97.2)
.20	.20	.125	97.9	98.0	98.0	96.9 (97.2)
.20	.60	.025	97.0	97.3	97.4	95.0 (94.8)
.20	.60	.075	96.2	96.2	96.5	95.0 (95.4)
.20	.60	.125	95.7	95.8	95.9	95.1 (96.5)

Table 2.8.1: Coverage of the upper limits

2.8.3 Estimates and confidence limits for $\mathbf{p_0}$ $\mathbf{(n_1 = 50)}$

Sample results		Classical		Bayesian non informative		Bayesian only $\alpha_1 = 5$	
n_1	c_0	$\widehat{p}_0$	p_0^u	b_0	b_0^u	b_0	b_0^u
50	4	.080	.174	.080	.171	.074	.159
50	5	.100	.199	.115	.195	.093	.182
50	6	.120	.223	.135	.219	.111	.204

Table 2.8.2: Estimates for a single sample check

Sample results					Classical		Bayesian non informative		Bayesian only $\alpha_{0\|0} = 5$	
n_1	n_2	c_0	c_{0+}	c_{10}	$\widehat{p}_0$	p_0^u	b_0	b_0^u	b_0	b_0^u
50	20	4	2	0	.080	.222	.093	.213	.087	.189
50	20	4	2	1	.131	.289	.132	.278	.117	.237
50	20	3	1	0	.060	.216	.071	.186	.065	.161
50	20	3	1	1	.106	.283	.109	.250	.094	.208
50	20	2	0	0	.040	.160	.049	.157	.044	.132
50	20	2	0	1	.088	.226	.085	.221	.071	.178
50	20	6	3	0	.120	.283	.136	.262	.129	.240
50	20	6	3	1	.172	.344	.176	.325	.161	.289
50	20	5	2	0	.100	.283	.114	.236	.108	.214
50	20	5	2	1	.150	.344	.153	.298	.138	.261
50	20	4	1	0	.080	.222	.092	.210	.086	.188
50	20	4	1	1	.128	.289	.130	.271	.116	.234
50	20	3	0	0	.060	.216	.070	.182	.065	.160
50	20	3	0	1	.107	.283	.107	.243	.093	.206

Table 2.8.3: Estimates for a double check with one error type

Sample results						Classical		Bayesian non informative		Bayesian only $\alpha_{0\|0}=5$	
n_1	n_2	c_0	c_{0+}	c_{10}	c_{01}	$\widehat{p}_0$	p_0^u	b_0	b_0^u	b_0	b_0^u
50	20	4	2	0	0	.080	.228	.081	.204	.075	.179
50	20	4	2	1	0	.131	.291	.122	.217	.107	.229
50	20	4	2	0	1	.040	.164	.043	.163	.040	.139
50	20	4	2	1	1	.091	.238	.085	.234	.073	.191
50	20	4	2	0	2	.000	.139	.000	.114	.000	.091
50	20	4	2	1	2	.051	.216	.046	.193	.038	.148
50	20	5	2	0	0	.100	.283	.096	.222	.091	.200
50	20	5	2	1	0	.150	.344	.137	.287	.124	.250
50	20	5	2	0	1	.050	.216	.051	.176	.049	.156
50	20	5	2	1	1	.100	.283	.094	.244	.085	.209
50	20	5	2	0	2	.000	.139	.000	.121	.000	.103
50	20	5	2	1	2	.050	.216	.049	.197	.044	.162
50	20	6	3	0	0	.120	.286	.122	.252	.116	.230
50	20	6	3	1	0	.178	.347	.164	.318	.150	.280
50	20	6	3	0	1	.080	.228	.085	.213	.080	.191
50	20	6	3	1	1	.132	.295	.128	.281	.115	.243
50	20	6	3	0	2	.040	.164	.044	.170	.041	.148
50	20	6	3	1	2	.092	.239	.089	.241	.078	.202
50	20	6	3	0	3	.000	.169	.000	.118	.000	.097
50	20	6	3	1	3	.052	.216	.047	.197	.040	.160

Table 2.8.4: Estimates for a double check with two error types

Chapter 3

Categorical data, multiple rounds

3.1 Introduction

Both the problem of missing data and the issue of misclassifications often occur in practice. Two main causes for missing observations are nonresponse and incomplete designs. While missing-by-design is due to incomplete designs and therefore is intentionally created by the experimenter, this is usually not true for nonresponse. Misclassifications occur in quality control where a checking device has to classify objects in ($r \geq 2$) categories, *e.g.* 'good' or 'bad'. Sometimes it is known that the checking device is fallible, but it might be too expensive or just impossible to procure a better one. In many situations both problems occur simultaneously: not only some observations are missing, but there may be misclassifications as well. A practical example of missing-by-design data with possible misclassifications is a repeated audit control.

In a repeated audit control one wants to draw conclusions about the fraction of elements in a population which belong to a certain category. In order to do this, an auditor classifies randomly sampled elements. However, misclassifications may occur, since the (usual) assumption that the auditor be infallible is dropped. To take these possible misclassifications into account, another fallible auditor checks a subsample of the already checked sample elements again. This procedure is repeated several times until the final k^{th} auditor, considered to be infallible, gives the true classification of some sample elements which already have been classi-

fied by all previous auditors. Conclusions about the population fractions have to be drawn based on the fallible and infallible audits. This kind of repeated audit control was introduced by Tenenbein (1970), who considered dichotomous data $(r = 2)$ and two audit rounds $(k = 2)$. This situation was further discussed in the previous chapter. Tenenbein (1972) extended the model to include categorical data $(r \geq 2)$.

Our Section 3.2 generalizes Chapter 2 into a general control system for categorical data ($r \geq 2$) with monotone missing observations obtained from $k \geq 2$ audit rounds. Subsamples for subsequent auditors are obtained by using either 'stratified' or 'random' sampling. Though these different sampling methods lead to different probability distributions, it is shown in Section 3.3 that the MLE's for the main parameters are identical. However, only in case of 'stratified' sampling do these MLE's appear to be unbiased. Special attention is paid to the frequently occurring situations in which the MLE's are undefined.

Since in auditing upper limits are very important, Section 3.4 considers three methods to obtain upper confidence limits for the population fractions; the Bayesian approach appears to be the most promising. Section 3.5 contains two practical applications, revisiting the Dutch social security case from the previous chapter. For $r = 2$ and $k = 3$ the calculation of Bayesian upper limits is presented in some detail. The final Section 3.6 contains the main conclusions and discusses our results.

3.2 A general model

3.2.1 Population model

Define the random variable I_0 as the true classification of a random sample element. The r possible classifications i_0 are denoted by $0, 1, \ldots, r-1$, while $p_{i_0} = \Pr(I_0 = i_0)$ denotes the population fraction of elements with true classification i_0.

A random element is classified by an auditor into one of the categories $0, 1, \ldots, r-1$, leading to the random variable I_1. Hence a correct classification only occurs if $I_1 = I_0$. To find possible misclassifications, the same element is categorized

once more, now by another auditor. This procedure is repeated, leading to classification I_j by auditor j, until the k^{th} auditor makes the final classification. Since this last auditor will be assumed to be an infallible expert, (s)he will always give the true classification: $I_k = I_0$.

The following notation will be used in the sequel to describe the different probabilities:

$$\begin{aligned} p_{i_0 i_1 \ldots i_j} &= Pr(I_0 = i_0, I_1 = i_1, \ldots, I_j = i_j), \quad j = 0, \ldots, k, \\ \pi_{i_1 i_2 \ldots i_j} &= Pr(I_1 = i_1, \ldots, I_j = i_j), \quad j = 1, \ldots, k. \end{aligned}$$

It seems unrealistic to assume that classifications of subsequent auditors are independent, even if previous classifications are hidden: indeed, previous classifications reveal the difficulty of correctly classifying a given element. For example, if many auditors judge an incorrect element to be correct, the error in the element probably is hard to detect. Hence we will need conditioning on previous classifications, to be denoted as follows:

$$\begin{aligned} p_{i_j|i_0 i_1 \ldots i_{j-1}} &= Pr(I_j = i_j | I_0 = i_0, \ldots, I_{j-1} = i_{j-1}), \quad j = 1, \ldots, k, \\ \pi_{i_j|i_1 \ldots i_{j-1}} &= Pr(I_j = i_j | I_1 = i_1, \ldots, I_{j-1} = i_{j-1}), \quad j = 2, \ldots, k. \end{aligned}$$

Since the last auditor is infallible ($I_k = I_0$), it follows $\pi_{i_1 i_2 \ldots i_k} = p_{i_0 i_1 \ldots i_k} = p_{i_0 i_1 \ldots i_{k-1}}$ for $i_k = i_0$. Other relations between the two sets of parameters are :

$$\begin{cases} \text{(a)} \quad \pi_{i_1 i_2 \ldots i_k} = p_{i_0} \cdot p_{i_1|i_0} \cdot p_{i_2|i_0 i_1} \cdot \ldots \cdot p_{i_{k-1}|i_0 i_1 \ldots i_{k-2}} \\ \text{(b)} \quad \pi_{i_1 i_2 \ldots i_k} = \pi_{i_1} \cdot \pi_{i_2|i_1} \cdot \pi_{i_3|i_1 i_2} \cdot \ldots \cdot \pi_{i_k|i_1 \ldots i_{k-1}} \\ \text{(c)} \quad p_{i_0} = p_{i_k} = \sum_{i_1 \ldots i_{k-1}} \pi_{i_1 i_2 \ldots i_k}. \end{cases} \tag{3.2.1}$$

Finally the following shorthand notations are introduced:

a :	one of the r^{j-1} possible classifications $i_1 i_2 \ldots i_{j-1}$ by the first $j-1$ auditors,
$p^{(0)}$:	row vector of r probabilities p_{i_0} ($i_0 = 0, 1, \ldots, r-1$),
$\pi_a^{(j)}$:	row vector of r probabilities π_{ai_j} ($i_j = 0, 1, \ldots, r-1$),
$\pi^{(j)}$:	$(r^{j-1} \times r)$ matrix with rows $\pi_a^{(j)}$,
$\pi_a^{(j\|j-1)}$:	row vector of r probabilities $\pi_{i_j\|a}$ ($i_j = 0, 1, \ldots, r-1$),
$\pi^{(j\|j-1)}$:	$(r^{j-1} \times r)$ matrix with rows $\pi_a^{(j\|j-1)}$,
$p_{i_0 a}^{(j\|j-1)}$:	row vector of r probabilities $p_{i_j\|i_0 a}$ ($i_j = 0, 1, \ldots, r-1$),
$p^{(j\|j-1)}$:	$(r^j \times r)$ matrix with rows $p_{i_0 a}^{(j\|j-1)}$.

The matrices are constructed with columnwise and rowwise decreasing classifications. These notations are illustrated below for $r = 2$.

$$\pi^{(1)} = \begin{pmatrix} \pi_1 & \pi_0 \end{pmatrix}, \qquad p^{(0)} = \begin{pmatrix} p_1 & p_0 \end{pmatrix},$$

$$\pi^{(2)} = \begin{pmatrix} \pi_1^{(2)} \\ \pi_0^{(2)} \end{pmatrix} = \begin{pmatrix} \pi_{11} & \pi_{10} \\ \pi_{01} & \pi_{00} \end{pmatrix}, \qquad \pi^{(2|1)} = \begin{pmatrix} \pi_{1|1} & \pi_{0|1} \\ \pi_{1|0} & \pi_{0|0} \end{pmatrix},$$

$$\pi^{(3)} = \begin{pmatrix} \pi_{11}^{(3)} \\ \pi_{10}^{(3)} \\ \pi_{01}^{(3)} \\ \pi_{00}^{(3)} \end{pmatrix} = \begin{pmatrix} \pi_{111} & \pi_{110} \\ \pi_{101} & \pi_{100} \\ \pi_{011} & \pi_{010} \\ \pi_{001} & \pi_{000} \end{pmatrix}, \qquad \pi^{(3|2)} = \begin{pmatrix} \pi_{1|11} & \pi_{0|11} \\ \pi_{1|10} & \pi_{0|10} \\ \pi_{1|01} & \pi_{0|01} \\ \pi_{1|00} & \pi_{0|00} \end{pmatrix},$$

$$p^{(2|1)} = \begin{pmatrix} p_{11}^{(2|1)} \\ p_{10}^{(2|1)} \\ p_{01}^{(2|1)} \\ p_{00}^{(2|1)} \end{pmatrix} = \begin{pmatrix} p_{1|11} & p_{0|11} \\ p_{1|10} & p_{0|10} \\ p_{1|01} & p_{0|01} \\ p_{1|00} & p_{0|00} \end{pmatrix}, \quad p^{(1|0)} = \begin{pmatrix} p_{1|1} & p_{0|1} \\ p_{1|0} & p_{0|0} \end{pmatrix}.$$

Consider a population which consists of incorrect ($i_0 = 0$) and correct elements ($i_0 = 1$). In order to draw conclusions about the population fraction of incorrect elements, a repeated audit control with three rounds is performed of which the last is infallible. Figure 3.2.1 gives an overview of the relevant probabilities; see also Figure 3.2.2.

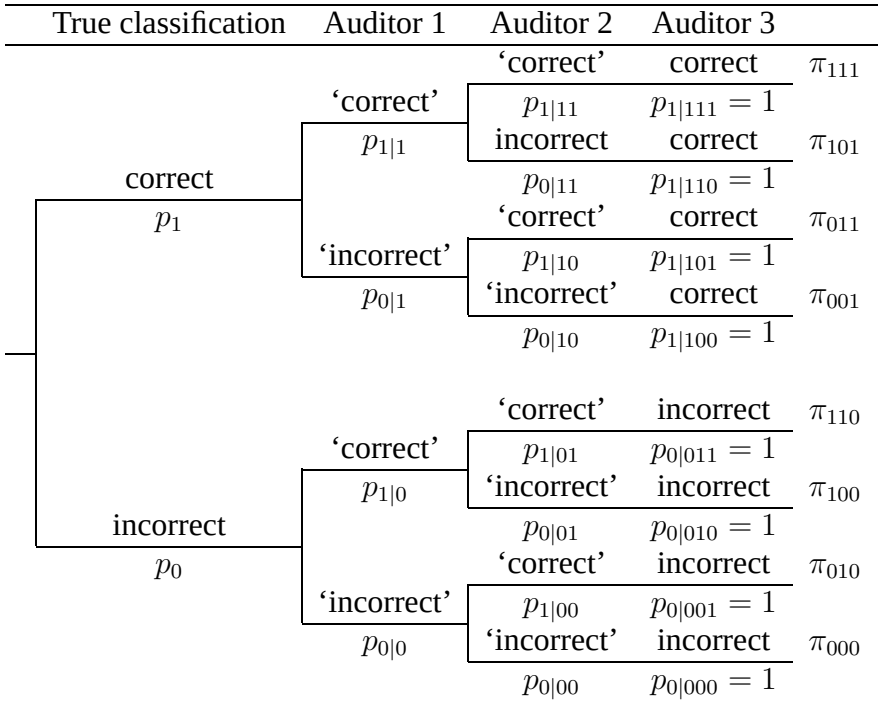


Figure 3.2.1: Classification probabilities $(r = 2, k = 3)$

3.2.2 Sample information

Auditor 1 classifies the elements of a random sample (drawn with replacement) of predetermined size n_1; a subsample of (possibly random) size $N_2 \leq n_1$ is checked again by auditor 2, and so on: auditor j checks $N_j \leq N_{j-1}$ elements $(j = 3, \ldots, k)$. Hence, N_k elements are classified by all auditors, $N_j - N_{j+1}$ elements by precisely the first j auditors. Such a pattern of observations is called a monotone missing data pattern; see Little and Rubin (2002). Note that here missing-by-design occurs.

Let C_a denote the number of elements classified by the first $j-1$ auditors as $a = i_1 \ldots i_{j-1}$. Of these, $N_a^{(j)} \leq C_a$ are observed by auditor j; the remainder $C_{a-} = C_a - N_a^{(j)}$ is not further investigated. The classification frequencies of auditor j are C_{ai_j} to be combined into the vector $C_a^{(j)}$. These r^{j-1} vectors can be collected into the matrix $C^{(j)}$, presenting all frequencies, observed by the first j

auditors. These notations agree with the notations for the parameters π. The k matrices $C^{(j)}$ summarize the complete sample information; compare Figure 3.2.2.

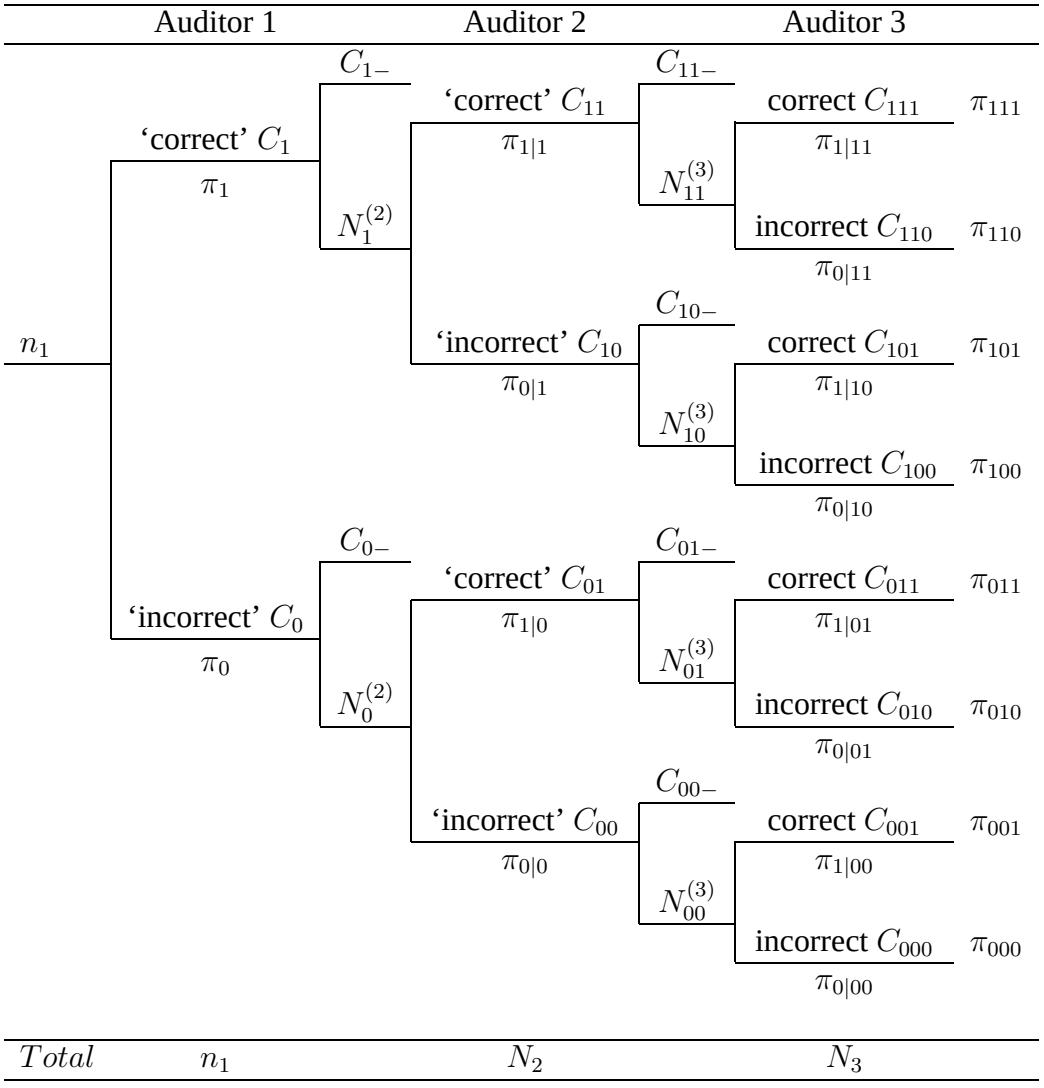


Figure 3.2.2: Classification frequencies and probabilities $(r = 2, k = 3)$

3.2.3 Sampling methods

An important aspect of a repeated audit control is the way in which it is decided which sample elements have to be checked again. In general, we allow the sample

sizes to depend on the preceding results. Two different sampling methods will be discussed here: stratified and random sampling. In case of stratified sampling, the sample size $N_a^{(j)}$ in round j from any given classification a is determined separately, while in random sampling only the total N_j over all these r^{j-1} classifications is prescribed. More precisely, let $\mathcal{C}^{(j)}$ denote the outcome space of $C^{(j)}$, while $f_a^{(j)}$ and g_j are given functions from $\mathcal{C}^{(1)} \cdot \mathcal{C}^{(2)} \cdot \ldots \cdot \mathcal{C}^{(j-1)}$ into $\mathbb{N} \cup \{0\}$ and $\mathbb{N}$, respectively, for all a and j. Then the two methods can be described as follows:

$$\begin{array}{ll} \text{stratified sampling:} & N_a^{(j)} = f_a^{(j)}(C^{(1)}, \ldots, C^{(j-1)}), \\ \text{random sampling:} & N_j = g_j(C^{(1)}, \ldots, C^{(j-1)}) \quad . \end{array}$$

Hence as soon as $C^{(j-1)}$ is known, the $N_a^{(j)}$ and N_j are given. Of course, the realization of the total sample size in round j also has to be positive for stratified sampling: $N_j = \sum_a N_a^{(j)} > 0$.

In most cases sample sizes will only depend on the previous round frequencies, so that $N_j = g_j(C^{(j-1)})$, *e.g.*; the simplest situation occurs when all the sample sizes are fixed predetermined numbers. This is the sampling method which is usually assumed in the existing literature on repeated audit controls.

3.3 Distributions and MLE's

3.3.1 Stratified sampling

All the following results are derived under the assumption of sampling with replacement. The convention that the multinomial distribution $M(0; .)$ is concentrated in 0 will be adopted.

Theorem 3.3.1. *In case of stratified sampling the joint sample distribution is characterized by the following multinomial distributions:*

$$\begin{cases} \mathcal{L}(C^{(1)}) = M(n_1; \pi^{(1)}), \\ \mathcal{L}(C_a^{(j)} | N_a^{(j)} = n_a^{(j)}) = M(n_a^{(j)}; \pi_a^{(j|j-1)}), \text{ for all } r^{j-1} \text{ possible } a, \quad j = 2, \ldots, k. \end{cases} \tag{3.3.1}$$

and the likelihood function $L(\pi^{(1)}, \pi_1^{(2|1)}, \ldots, \pi_a^{(k|k-1)}; c^{(1)}, \ldots, c^{(k)})$ *is obtained by multiplying all probabilities corresponding with the* $(1-r^k)/(1-r)$ *multinomials in (3.3.1).*

Proof. Equation (3.3.1) is obvious. Further, because the $f_a^{(j)}$ are given functions,

$$\mathcal{L}(C_a^{(j)}|C^{(1)}, \ldots, C^{(j-1)}) = \mathcal{L}(C_a^{(j)}|N_a^{(j)})$$

holds for all a and j, while these distributions are conditionally independent for different a. This implies the second statement. □

The corresponding log-likelihood follows at once:

$$log(L(\pi^{(1)}, \pi_1^{(2|1)}, \ldots, \pi_a^{(k|k-1)}; c^{(1)}, \ldots, c^{(k)})) = \sum_{i_1} c_{i_1} \log \pi_{i_1} + \sum_{j=2}^{k} \sum_{ai_j} c_{ai_j} \log \pi_{i_j|a} \tag{3.3.2}$$

as well as the MLE's for all parameters involved:

$$\begin{cases} \widehat{\Pi}^{(1)} & = \dfrac{C^{(1)}}{n_1} \\ \widehat{\Pi}_a^{(j|j-1)} & = \dfrac{C_a^{(j)}}{N_a^{(j)}}, \quad \text{for all } r^{j-1} \text{ possible } a, \quad j = 2, \ldots, k. \end{cases} \tag{3.3.3}$$

These MLE's are the regular MLE's for a k-way contingency table with $k-1$ supplementary marginal tables with MAR (missing at random) multinominal data (see Little and Rubin (2002) for more details). Since the parameters of interest p_{i_0} are functions of $(\pi_{i_1}, \pi_{i_j|a})$ (see (3.2.1)), the MLE's for p_{i_0} are functions of the MLE's in (3.3.3):

$$\widehat{P}_{i_0} = \widehat{P}_{i_k} = \sum_{i_1 \ldots i_{k-1}} \widehat{\Pi}_{i_1 i_2 \ldots i_k} = \sum_{i_1 \ldots i_{k-1}} \widehat{\Pi}_{i_1} \cdot \widehat{\Pi}_{i_2|i_1} \cdot \ldots \cdot \widehat{\Pi}_{i_k|i_1 \ldots i_{k-1}}. \tag{3.3.4}$$

However, the MLE's for the conditional classification probabilities $\pi_{i_j|a}$ are not defined when $N_a^{(j)} = 0$. This is asymptotically irrelevant but highly relevant in practice! Although the probability of undefined ML estimates tends asymptotically to zero, practical repeated audit controls usually have small final sample

sizes due to the high costs of the last auditor. Undefined MLE's are (in general) frequently occurring and it is important to have a good estimation procedure which can handle these situations. Section 3.3.3 examines possible procedures for undefined MLE's more closely.

Note that the auditors' error probabilities can be derived from (3.2.1), (3.3.3) and (3.3.4) as well; *e.g.*

$$\widehat{P}_{i_1|i_0} = \widehat{P}_{i_1|i_k} = \frac{\widehat{P}_{i_1 i_k}}{\widehat{P}_{i_k}} = \frac{\sum\limits_{i_2 \dots i_{k-1}} \widehat{\Pi}_{i_1 i_2 \dots i_k}}{\sum\limits_{i_1 \dots i_{k-1}} \widehat{\Pi}_{i_1 i_2 \dots i_k}} = \frac{\sum\limits_{i_2 \dots i_{k-1}} \widehat{\Pi}_{i_1} \cdot \widehat{\Pi}_{i_2|i_1} \cdot \ldots \cdot \widehat{\Pi}_{i_k|i_1 \dots i_{k-1}}}{\sum\limits_{i_1 \dots i_{k-1}} \widehat{\Pi}_{i_1} \cdot \widehat{\Pi}_{i_2|i_1} \cdot \ldots \cdot \widehat{\Pi}_{i_k|i_1 \dots i_{k-1}}}.$$

3.3.2 Random sampling

Although the $N_a^{(j)}$ are deterministic conditionally on the previous classifications in the case of stratified sampling, this is not true for random sampling and the characteristic distributions differ for the two sampling methods. Let $N^{(j)}$ denote the vector of all r^{j-1} scalars $N_a^{(j)}$.

Theorem 3.3.2. *In case of random sampling the joint sample distribution is characterized by the following multinomial distributions:*

$$\begin{cases} \mathcal{L}(C^{(1)}) = M(n_1; \pi^{(1)}) \\ \mathcal{L}(N^{(j)} | C^{(j-1)} = c^{(j-1)}, N_j = n_j) = M(n_j; \dfrac{vec(c^{(j-1)})}{n_{j-1}}), \qquad j = 2, \dots, k \\ \mathcal{L}(C_a^{(j)} | N_a^{(j)} = n_a^{(j)}) = M(n_a^{(j)}; \pi_a^{(j|j-1)}), \text{ for all } r^{j-1} \text{ possible } a, \ j = 2, \dots, k. \end{cases} \tag{3.3.5}$$

and the likelihood inference is the same as for stratified sampling.

Proof. The conditional multinomial distribution functions (3.3.5) are again straightforward. The likelihood is now acquired by multiplying all the $(1-r^k)/(1-r) + k - 1$ conditionally independent multinomial distributions:

$$\begin{aligned} &L(\pi^{(1)}, \pi_1^{(2|1)}, \dots, \pi_a^{(k|k-1)}; c^{(1)}, \dots, c^{(k)}, n^{(2)}, \dots, n^{(k)}) \\ &\quad = \mathcal{L}(C^{(1)}) \mathcal{L}(N^{(2)} | C^{(1)}, n_2) \mathcal{L}(C_0^{(2)} | n_0^{(2)}) \mathcal{L}(C_1^{(2)} | n_1^{(2)}) \cdot \ldots \cdot \mathcal{L}(C_a^{(k)} | n_a^{(k)}). \end{aligned}$$

The conditional distribution functions for the classification quantities $C^{(1)}$ and $C_a^{(j)}$ are identical for random and stratified subsampling. Therefore the likelihood functions of the two sampling methods differ only by the additional conditional distribution functions of the sample sizes $N^{(j)}$ in case of random sampling. Since these distribution functions do not depend on the parameters, the distributions of the $N^{(j)}$ can be ignored for likelihood inferences about the parameters: $C^{(1)}$ and $C_a^{(j)}$ are sufficient for π_{i_1} and $\pi_{i_j|a}$, respectively. □

3.3.3 Undefined MLE's

Though the MLE's have nice asymptotic properties and are logically interpretable, a major drawback is that they will be frequently undefined in practice (depending on the sampling method). The MLE's for the population fractions are undefined when auditor j does not classify at least one sample element of each previously occurring classification pattern, *i.e.* $n_a^{(j)} = 0$ while $c_a > 0$. The situation $n_a^{(j)} = 0$ can be divided into structural zeros and unstructural zeros (see Bishop *et al.* (1975)). Unstructural zeros are caused by chance while structural zeros are caused by *a priori* model restrictions such as $\pi_a = 0$. In this chapter we extend this last definition to include the situation $n_a^{(j)} = 0$ when $c_a > 0$, where the elements with previous classification a are intentionally excluded from the j^{th} sample ($N_a^{(j)} = f_a^{(j)}(C^{(1)}, \ldots, C^{(j-1)}) = 0$) because another check would not provide additional information.

Consider for example a population which consists of correct ($i_0 = 1$) and incorrect elements ($i_0 = 0$). A repeated audit control takes place with only one fallible auditor $(k = 2)$. The fallible auditor is *a priori* known never to misclassify correct elements $(p_{1|1} = 1)$ but (s)he might make mistakes with incorrect elements. As a consequence an element which the first auditor classifies as incorrect is per definition incorrect. An additional check of such an element does not provide extra information and is therefore useless. A logical choice is $N_0^{(2)} = 0$. Though $\widehat{\Pi}_{1|0}$ is now undefined according to (3.3.3), this is not a problem since it is *a priori* known that $\pi_{1|0} = 0$.

In general, structural zeros do not cause problems because they are caused

themselves by model assumptions about the parameters. Unstructural zeros, however, are the cause of some problems. Fortunately, unstructural zeros can be avoided completely by using a specific kind of stratified sampling: stratified sampling with $N_a^{(j)} > 0$ when $c_a > 0$. In these cases the MLE's for p_{i_0} are always uniquely defined and are even unbiased.

Theorem 3.3.3. $E\{\widehat{P}_{i_0}\} = p_{i_0}$ *if* $N_a^{(j)} > 0$ *when* $C_a > 0$.

Proof. If $N_a^{(j)} > 0$ when $C_a > 0$, the MLE's $\widehat{\Pi}_{i_j|a}$ in $(3.3.3)$ can still be undefined. However, the preceding factor $\widehat{\Pi}_{i_{j-1}|i_1 \ldots i_{j-2}}$ in (3.3.4) is per definition 0 when $N_a^{(j)} = 0$. As a consequence, the corresponding term $\widehat{\Pi}_{i_1 \ldots i_k}$ of $\widehat{P}_{i_0}$ in (3.3.4) is zero. So the MLE's $\widehat{P}_{i_0}$ are defined, even in case of undefined MLE's for the conditional classification probabilities. From the relations

$$\begin{aligned} E\{\widehat{\Pi}_{i_1 i_2 \ldots i_j}\} &= E\{\widehat{\Pi}_a \cdot \widehat{\Pi}_{i_j|a}\} = E\{\widehat{\Pi}_a\} E\{\frac{C_{a i_j}}{N_a^{(j)}} | N_a^{(j)}\} \\ &= E\{\widehat{\Pi}_a\} \cdot \pi_{i_j|a} = E\{\widehat{\Pi}_{i_1 \ldots i_{j-1}}\} \cdot \pi_{i_j|i_1 \ldots i_{j-1}}, \end{aligned}$$

it follows by repeated application that $E\{\widehat{\Pi}_{i_1 i_2 \ldots i_j}\} = \pi_{i_1 i_2 \ldots i_j}$. In combination with (3.2.1), this gives

$$E\{\widehat{P}_{i_0}\} = \sum_{i_1 \ldots i_{k-1}} E\{\widehat{\Pi}_{i_1 i_2 \ldots i_k}\} = \sum_{i_1 \ldots i_{k-1}} \pi_{i_1 i_2 \ldots i_k} = p_{i_0},$$

which completes the proof. □

A disadvantage of this kind of stratified sampling is that the required final sample size can be quite large since the last sample has to include at least one element of all previous realized classifications. This could be an argument to apply a different sampling method which could still lead to unstructural zeros. Section 3.5.1 shows that a procedure for handling situations with undefined MLE's is indeed important.

3.4 Upper limits

3.4.1 Classical; finite samples

For a standard audit with an infallible auditor ($k = 1$) and dichotomous data ($r = 2$) the upper $(1-\alpha)-$confidence limit for p_{i_0}, denoted by $p^u_{i_0}$ is the regular binomial confidence limit

$$p^u_{i_0} = \underset{p_{i_0}}{max} \left\{ p_{i_0} : Pr(\widehat{P}_{i_0} \leq \widehat{p}_{i_0} | p_{i_0}) \geq \alpha \right\}. \tag{3.4.1}$$

The generalization for $r = 2$ and $k = 2$ is given in (2.4.1), which we repeat here for convenience:

$$p^u_0 = \underset{p_{i_0}}{max} \left\{ p_0, p_{1|0}, p_{0|1} : Pr(\widehat{P}_0 \leq \widehat{p}_0 | p_0, p_{1|0}, p_{0|1}) \geq \alpha \right\}. \tag{3.4.2}$$

To determine this upper limit, the maximum $p^u_0 | p_{1|0}, p_{0|1}$ of (3.4.2) for fixed $p_{1|0}$ and $p_{0|1}$ has to be calculated for all possible values of the nuisance parameters $p_{1|0}$ and $p_{0|1}$. Subsequently, p^u_0 is determined as the maximum of all $p^u_0 | p_{1|0}, p_{0|1}$. Compare Section 2.4.

It is straightforward to generalize (3.4.2) for $r \geq 2$ and $k \geq 2$:

$$p^u_{i_0} = \underset{p_{i_0}}{max} \left\{ p_{i_0}, p^{(j|j-1)} : Pr(\widehat{P}_{i_0} \leq \widehat{p}_{i_0} | p_{i_0}, p^{(j|j-1)},\ j = 1, \ldots, k-1) \geq \alpha \right\}.$$

The determination of $p^u_{i_0}$ runs as in the case $r = 2$ and $k = 2$.

A disadvantage of this method is the worst case approach: while determining the upper limit all situations (*i.e.* all values of the nuisance parameters) are considered and the most unfavorable one is chosen. All possible situations also include the situation in which each fallible auditor deliberately classifies all elements in the same category regardless of the true and previous classifications, *i.e.* for $j = 1, \ldots, k-1$ the elements of $p^{(j|j-1)}$ consist solely of zeros and ones. As a consequence all elements will be classified in exactly the same way by the first $k-1$ auditors: $i^*_1, \ldots, i^*_{k-1}$. In this case the MLE's in (3.3.4) reduce to

$$\widehat{P}_{i_0} = \quad \widehat{P}_{i_k} = \widehat{\Pi}_{i^*_1 i^*_2 \ldots i^*_{k-1} i_k} = \quad \widehat{\Pi}_{i_k | i^*_1 \ldots i^*_{k-1}} = \frac{C_{i^*_1 \ldots i^*_{k-1} i_k}}{N_k}.$$

The latter is just the estimator in case of an ordinary audit with only an infallible auditor who checks $(n_1 =)N_k$ elements. So $p^u_{i_0}|p^{(j|j-1)}$ is solely based on the classifications by the last infallible auditor and the fallible classifications are disregarded completely. Therefore it coincides with the upper limit (3.4.1) of a standard audit by an infallible auditor who checks N_k elements. As a consequence $p^u_{i_0}$, which is the maximum of all $p^u_{i_0}|p^{(j|j-1)}$ will be at least as high as (3.4.1) and the repeated audit control is in this sense useless: the fallible classifications cost money but do not provide more accurate estimates.

So although the described method enables us to find confidence limits for finite samples, these confidence limits will be very high since the - often unlikely - worst case is taken to be reality. This conclusion is in line with the results of the previous chapter.

3.4.2 Classical; limit distributions

A widely applied approach to construct confidence intervals is based on the limit distribution of the MLE's.

Theorem 3.4.1. *Under the assumption* $N^{(j)}_a/n_1 \xrightarrow{P} b^{(j)}_a$ *if* $n_1 \to \infty$*, with* $b^{(j)}_a$ *a constant depending on* a*,*

$$\sqrt{n_1}(\widehat{P}_{i_0} - p_{i_0}) \xrightarrow{\mathcal{L}} N(0, \sigma^2_{i_0}), \tag{3.4.3}$$

with

$$\sigma^2_{i_0} = \sigma^2_{i_k} = \sum_{i_1 \ldots i_{k-1}} Var(\widehat{\Pi}_{i_1 \ldots i_{k-1} i_k}) + \sum_{i_1 \ldots i_{k-1} \neq i'_1 \ldots i'_{k-1}} Cov(\widehat{\Pi}_{i_1 \ldots i_{k-1} i_k}, \widehat{\Pi}_{i'_1 \ldots i'_{k-1} i_k}).$$

Define $w = min\{j : i_j \neq i'_j$ *for* $i_1 \ldots i_k$ *and* $i'_1 \ldots i'_k\}$ *then*

$$Cov(\widehat{\Pi}_{i_1 \ldots i_k}, \widehat{\Pi}_{i'_1 \ldots i'_k}) =$$

$$\begin{cases} -\pi_{i_1\ldots i_k}\pi_{i_1'\ldots i_k'} & \text{if } w = 1 \\ \dfrac{\pi_{i_1\ldots i_k}}{\pi_{i_1}}\dfrac{\pi_{i_1'\ldots i_k'}}{\pi_{i_1}}\pi_{i_1}(1-\pi_{i_1}) + \\ \sum_{j=2}^{w-1}\dfrac{\pi_{i_1\ldots i_k}}{\pi_{i_j|a}}\dfrac{\pi_{i_1'\ldots i_k'}}{\pi_{i_j|a}}\dfrac{\pi_{i_j|a}(1-\pi_{i_j|a})}{b_a^{(j)}} - \dfrac{\pi_{i_1\ldots i_k}\pi_{i_1'\ldots i_k'}}{b^{(w)}_{i_1\ldots i_{w-1}}} & \text{if } 1 < w \leq k \\ \left(\dfrac{\pi_{i_1\ldots i_k}}{\pi_{i_1}}\right)^2\pi_{i_1}(1-\pi_{i_1}) + \sum_{j=2}^{k}\left(\dfrac{\pi_{i_1\ldots i_k}}{\pi_{i_j|a}}\right)^2\dfrac{\pi_{i_j|a}(1-\pi_{i_j|a})}{b_a^{(j)}} & \text{else.} \end{cases}$$

Proof. See Appendix 3.7.1. □

Now the standard techniques can be applied to construct confidence intervals. Tenenbein (1970), Tenenbein (1971), Tenenbein (1972) used the variance of the limit distribution $\sigma_{i_0}^2$ as a measure of accuracy of the repeated audit control. However, as mentioned before, asymptotics are often not relevant for these types of controls.

Neither of the two methods for constructing confidence intervals which are discussed so far, appears to be very useful. Therefore, we consider the Bayesian approach as well.

3.4.3 Bayesian

In the Bayesian approach for monotone missing multinomial data, prior distributions can be chosen for either the set of parameters π (all $\pi^{(j)}$ and $\pi^{(j|j-1)}$), or all parameters p ($p^{(0)}$ and $p^{(j|j-1)}$); of course these parameters now are seen as random variables (which will be denoted by the corresponding upper cases, *e.g.* the two sets of parameters will be denoted as Π and P). The first choice is the simplest; in that case independent Dirichlet distributions often are taken as priors. Combined with the data, they lead to a simultaneous posterior distribution for the variables Π which is the product of independent Dirichlet distributions (see *e.g.* Schafer (1997)). Our parameter of interest P_0 is a known function of Π and its marginal posterior distribution can be straightforward determined by means of simulation from the posterior distribution of Π. The mode and $(1-\alpha)$-quantile

from the marginal distribution can be taken as point estimate and upper limit, respectively.

However, since our parameter of interest is p_0 and our model is originally formulated in terms of p, a more logical choice is to formulate priors for P in stead of Π. Moreover, independent (Dirichlet) priors for P seem reasonable since the quality of the population and the different auditors are likely not to depend on each other. This argumentation for independence does not hold for Π. Therefore the product of the following independent Dirichlet distributions is taken as prior:

$$\begin{cases} \mathcal{L}(P^{(0)}) = D(\alpha_{r-1}, \alpha_{r-2}, \ldots, \alpha_0) \\ \mathcal{L}(P_{i_0a}^{(j|j-1)}) = D(\alpha_{r-1|i_0a}, \alpha_{r-2|i_0a}, \ldots, \alpha_{0|i_0a},), \quad \forall a, \forall i_0, \forall j. \end{cases} \tag{3.4.4}$$

Since the data are missing at random (see Rubin (1976)), distribution (3.3.1) suffices for the Bayesian inference, regardless whether random or stratified sampling is applied. The simultaneous posterior distribution of P is the product of (3.4.4) and (3.3.1). The marginal posterior distribution of P_0 is obtained by integration. This is analytically rather complicated but can also be done by means of simulation or numerical integration, as in Chapter 2 for $r = 2$ and $k = 2$. However, instead of integrating the simultaneous posterior distribution, it is also possible to determine the marginal posterior distribution by means of the data augmentation algorithm of Tanner and Wong (1987).

Data augmentation is an iterative method of simulating the posterior distribution for missing data problems. The basic idea is that the required posterior distribution would be straightforward to determine if there were no missing observations. For our model it is easy to verify that P would have the following Dirichlet posteriors in case of the Dirichlet priors (3.4.4) and complete data:

$$\begin{cases} \mathcal{L}(P^{(0)}|\text{data}) = D(\alpha^{(0)} + c^{[k]}) \\ \mathcal{L}(P_{i_0a}^{(j|j-1)}|\text{data}) = D(\alpha_{i_0a}^{(j|j-1)} + c_{ai_0}^{[j]}), \quad \forall a, \forall i_0, \forall j, \end{cases} \tag{3.4.5}$$

where

$\alpha^{(0)}$	:	vector of exponents α_{i_0} corresponding with the vector $P^{(0)}$,
$\alpha_{i_0a}^{(j|j-1)}$	:	vector of exponents $\alpha_{i_j|i_0a}$ corresponding with the vector $P_{i_0a}^{(j|j-1)}$,
$c_{ai_0}^{[j]}$	:	vector of the numbers $c_{ai_j+\cdots+i_0}$ of classifications ai_j by the first j auditors, i_0 by the last (and any classification by auditors $j+1, \ldots, k-1$).

Each iteration of the data augmentation procedure consists of an imputation step and the posterior step. Start with an initial draw of the parameters from an approximation to the posterior distribution. In the imputation step the missing data are drawn from the appropriate distribution (with the drawn parameters) to get a (simulated) complete dataset. In the subsequent posterior step the parameters are drawn from the complete data posterior. Given the newly drawn parameters the imputation step is again executed, *et cetera*.

For our model, the imputation step consists of drawing the missing observations from a multinomial distribution:

$$\begin{cases} \mathcal{L}(\widehat{C}_{i_1}^{(2)}) = M(c_{i_1}; \pi_{i_1}^{(2|1)}), & \forall i_1 \\ \mathcal{L}(\widehat{C}_a^{(j)}) = M(c_{a-} + \widehat{c}_a; \pi_a^{(j|j-1)}), & \forall a, \quad j = 3, \ldots, k. \end{cases} \tag{3.4.6}$$

The p are drawn from posterior distributions which are similar to (3.4.5):

$$\begin{cases} \mathcal{L}(P^{(0)}|\text{(simulated) data}) = D(\alpha^{(0)} + c^{[k]} + \widehat{c}^{[k]}) \\ \mathcal{L}(P_{i_0 a}^{(j|j-1)}|\text{(simulated) data}) = D(\alpha_{i_0 a}^{(j|j-1)} + c_{a i_0}^{[j]} + \widehat{c}_{a i_0}^{[j]}), \quad \forall a, \forall i_0, \forall j. \end{cases} \tag{3.4.7}$$

The π, which are required for the subsequent imputation step, can now be determined from (3.2.1). In Section 3.5.2, the data augmentation algorithm will be applied to an example with $r = 2$ and $k = 3$.

3.5 Applications

3.5.1 Case *r=2, k=2*

A population consists of correct ($i_0 = 1$) and incorrect ($i_0 = 0$) elements. In order to estimate p_0, a repeated audit control is performed by two auditors. Random sampling is applied with n_2 being a fixed number: $N_2(C_1, C_0) = n_2$. There are no prior assumptions about the quality of the first auditor, *i.e.* about the misclassification probabilities. The characteristic sample distributions (3.3.5) are reduced

to:

$$\begin{cases} \mathcal{L}(C_1, C_0) = M(n_1; \pi_1, \pi_0) \\ \mathcal{L}(N^{(2)}|C_1 = c_1, C_0 = c_0) = M(n_2; c_1/n_1, c_0/n_1) \\ \mathcal{L}(C_{11}, C_{10}|N_1^{(2)} = n_1^{(2)}) = M(n_1^{(2)}; \pi_{1|1}, \pi_{0|1}) \\ \mathcal{L}(C_{01}, C_{00}|N_0^{(2)} = n_0^{(2)}) = M(n_0^{(2)}; \pi_{1|0}, \pi_{0|0}), \end{cases}$$

- compare (2.2.4) - and the MLE's (3.3.4) follow.

Both Tenenbein (1970), Moors (1999) and Barnett *et al*. (2001) derived these MLE's. Tenenbein (1970) noted that the MLE for p_0 is undefined when either $N_0^{(2)}$ or $N_1^{(2)}$ equals 0, but he concluded that the probability of this occurring is quite small unless n_2 is small and π_1 or π_0 is close to zero. However, these cases are of importance for calculating upper confidence limits. Moors (1999) derived the MLE's independently from Tenenbein (1970) and paid special attention to the cases of undefined MLE's. To determine the MLE's in these cases with only 'correct' or 'incorrect' sample records in the second round, he made the extra assumption $p_{1|0} = 1 - p_{0|1}$. This resulted in estimator (2.3.2), which in the present notation reads:

$$\widehat{P}_0 = \begin{cases} \dfrac{C_{10}}{N_1^{(2)}} & \text{if } N_0^{(2)} = 0 \\ \dfrac{C_0}{n_1}\dfrac{C_{00}}{N_0^{(2)}} + \dfrac{C_1}{n_1}\dfrac{C_{10}}{N_1^{(2)}} & \text{if } 0 < N_0^{(2)} < n_2 \\ \dfrac{C_{00}}{N_0^{(2)}} & \text{if } N_0^{(2)} = n_2. \end{cases} \tag{3.5.1}$$

The main expression consists of two terms which have a logical interpretation. The first term is the fraction of elements which are classified as 'incorrect' by the first auditor times the estimated probability that they are indeed incorrect. The second term is the fraction of elements which are classified as 'correct' by the first auditor times the estimated probability that they are actually incorrect. If either $N_0^{(2)}$ or $N_1^{(2)}$ equals 0, all information of the fallible auditor is discarded.

Table 3.5.1 contains the numerical data (in the present notation) of the CTSV example of Chapter 2 (compare Table 2.4.1).

	Total	Single checked sample	Double checked sample Second auditor		
First auditor			Total	correct	incorrect
'correct'	$c_1 = 484$	$c_{1-} = 433$	$n_1^{(2)} = 51$	$c_{11} = 50$	$c_{10} = 1$
'incorrect'	$c_0 = 16$	$c_{0-} = 14$	$n_0^{(2)} = 2$	$c_{01} = 0$	$c_{00} = 2$
Total	$n_1 = 500$	$n_1 - n_2 = 447$	$n_2 = 53$	$c_{+1} = 50$	$c_{+0} = 3$

Table 3.5.1: CTSV example

For this practical example, estimator (3.5.1) leads to a point estimate of 0.0510; the 95% upper confidence level was 0.121 - obtained from (3.4.2). In the next section, this CTSV example will be used again.

The major disadvantage of Moors' estimator $\widehat{P}_0$ is that it does not coincide with the MLE for the reduced models. In a reduced model, one misclassification probability, either $p_{1|0}$ or $p_{0|1}$, is *a priori* set to zero. It can be shown that Moors' estimator does not coincide with the MLE of the two reduced models if either $N_0^{(2)}$ or $N_1^{(2)}$ equals 0. Therefore, a slightly modified estimator is proposed:

$$\widehat{P}_0^* = \begin{cases} \dfrac{C_0}{n_1} + \dfrac{C_1}{n_1}\dfrac{C_{10}}{N_1^{(2)}} & \text{if } N_0^{(2)} = 0 \\ \dfrac{C_0}{n_1}\dfrac{C_{00}}{N_0^{(2)}} + \dfrac{C_1}{n_1}\dfrac{C_{10}}{N_1^{(2)}} & \text{if } 0 < N_0^{(2)} < n_2 \\ \dfrac{C_0}{n_1}\dfrac{C_{00}}{N_0^{(2)}} & \text{if } N_0^{(2)} = n_2. \end{cases} \tag{3.5.2}$$

This is the only estimator which coincides with the MLE of the reduced models. In order to see whether the differences between (3.5.1) and (3.5.2) are relevant, a comparison is made based on the bias. By taking conditional expectations (see Appendix 3.7.2) it follows:

$$\begin{aligned} Bias(\widehat{P}_0) &= (1 - \frac{n_2}{n_1})(\pi_1^{n_2}(\pi_{0|1} - \pi_{00} - \pi_{10}) + \pi_0^{n_2}(\pi_{0|0} - \pi_{00} - \pi_{10})), \\ Bias(\widehat{P}_0^*) &= (1 - \frac{n_2}{n_1})(\pi_1^{n_2}\pi_{01} - \pi_0^{n_2}\pi_{10}). \end{aligned}$$

The bias of both estimators depends on the classification probabilities and the sample sizes. The bias is reduced by increasing n_2 or decreasing n_2/n_1. This means that the bias is smaller if more infallible information is acquired or if the

fraction of fallible information decreases. The bias of $\widehat{P}_0^*$ decreases when the first auditor is more accurate; it is even unbiased in the case of an infallible first auditor. The latter is not true for $\widehat{P}_0$. Figure 3.5.1 shows that the difference between the estimators can be quite substantial.

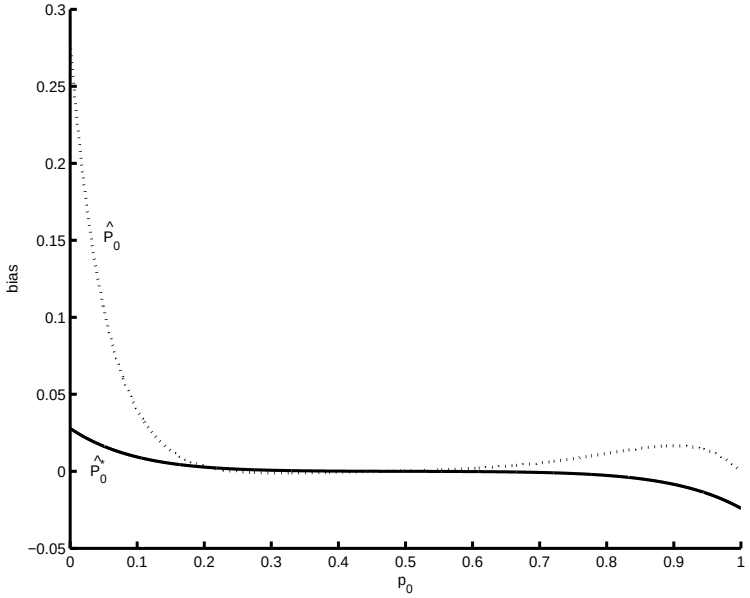


Figure 3.5.1: Bias of $\widehat{P}_0$ and $\widehat{P}_0^*$

This graph shows the bias of estimators (3.5.1) and (3.5.2) for $n_1 = 50$, $n_2 = 10$, $p_{1|0} = 0.05$ and $p_{0|1} = 0.10$. In particular for low values of p_0, use of the modified estimator $\widehat{P}_0^*$ leads to a generally much smaller bias.

For $r = 2$ and $k = 2$ an analytical expression for the posterior distribution can be given; analysis and results are presented in Chapter 2. Application of the data augmentation procedure leads to identical results.

3.5.2 Case *r=2, k=3*

In the previous subsection, we discussed the CTSV example in which a repeated audit control with two rounds was applied. However, the CTSV also applied repeated audit controls with three rounds. In the first two rounds (fallible) internal

auditors of the six companies classified the (sub)sampled security payments as correct or incorrect. In the third and final round an auditor of the CTSV checked a subsample of the twice checked payments. Again, the auditor of the CTSV is considered to be flawless.

Since we do not have access to data of the three rounds, we use the previously analysed data of the repeated audit control with two rounds (see Table 3.5.1), but extend it with fictitious data for the third round.

In this third check, the infallible expert once more classifies a subsample of size $n_3 = 20$ of the 53 double checked payments; all payments considered incorrect by at least one of the two internal auditors are included. This results in the following (stratified) sample sizes:

$$n_3 = 20, \quad n_{00}^{(3)} = c_{00} = 2, \quad n_{10}^{(3)} = c_{10} = 1, \quad n_{01}^{(3)} = c_{01} = 0, \quad n_{11}^{(3)} = 17.$$

For the outcomes of this third check, the four different possibilities in Table 3.5.2 are considered.

Possibility 1		Possibility 2	
correct	incorrect	correct	incorrect
$c_{111} = 17$	$c_{110} = 0$	$c_{111} = 16$	$c_{110} = 1$
$c_{101} = 0$	$c_{100} = 1$	$c_{101} = 0$	$c_{100} = 1$
$c_{011} = 0$	$c_{010} = 0$	$c_{011} = 0$	$c_{010} = 0$
$c_{001} = 0$	$c_{000} = 2$	$c_{001} = 0$	$c_{000} = 2$
$c_{++1} = 17$	$c_{++0} = 3$	$c_{++1} = 16$	$c_{++0} = 4$

Possibility 3		Possibility 4	
correct	incorrect	correct	incorrect
$c_{111} = 17$	$c_{110} = 0$	$c_{111} = 17$	$c_{110} = 0$
$c_{101} = 1$	$c_{100} = 0$	$c_{101} = 0$	$c_{100} = 1$
$c_{011} = 0$	$c_{010} = 0$	$c_{011} = 0$	$c_{010} = 0$
$c_{001} = 0$	$c_{000} = 2$	$c_{001} = 1$	$c_{000} = 1$
$c_{++1} = 18$	$c_{++0} = 2$	$c_{++1} = 18$	$c_{++0} = 2$

Table 3.5.2: Fictitious data third round

In Possibility 1, the expert fully agrees with the second auditor. In Possibility 2, one error is missed by both fallible auditors; further the expert fully agrees with

the second auditor. In the third option, the expert fully agrees with the first auditor implying that the second auditor missed one incorrect payment. In Possibility 4, the expert finds that one error is made up by both auditors; further findings are in agreement with the second auditor.

The general MLE (3.3.4) reduces in this case ($r = 2, k = 3$) to:

$$\widehat{P}_0 = \sum_{i=0}^{1}\sum_{j=0}^{1} \frac{C_i}{n_1} \cdot \frac{C_{ij}}{N_i^{(2)}} \cdot \frac{C_{ij0}}{N_{ij}^{(3)}}.$$

This estimator is defined for all possibilities of the numerical example. The point estimates for the π's and p's - if defined - are shown in Table 3.5.3.

	$\widehat{\pi}_0$	$\widehat{\pi}_{0\|0}$	$\widehat{\pi}_{0\|1}$	$\widehat{\pi}_{0\|00}$	$\widehat{\pi}_{0\|01}$	$\widehat{\pi}_{0\|10}$	$\widehat{\pi}_{0\|11}$
Possibility 1	0.0320	1.0000	0.0196	1.0000	-	1.0000	0.0000
Possibility 2	0.0320	1.0000	0.0196	1.0000	-	1.0000	0.0588
Possibility 3	0.0320	1.0000	0.0196	1.0000	-	0.0000	0.0000
Possibility 4	0.0320	1.0000	0.0196	0.5000	-	1.0000	0.0000

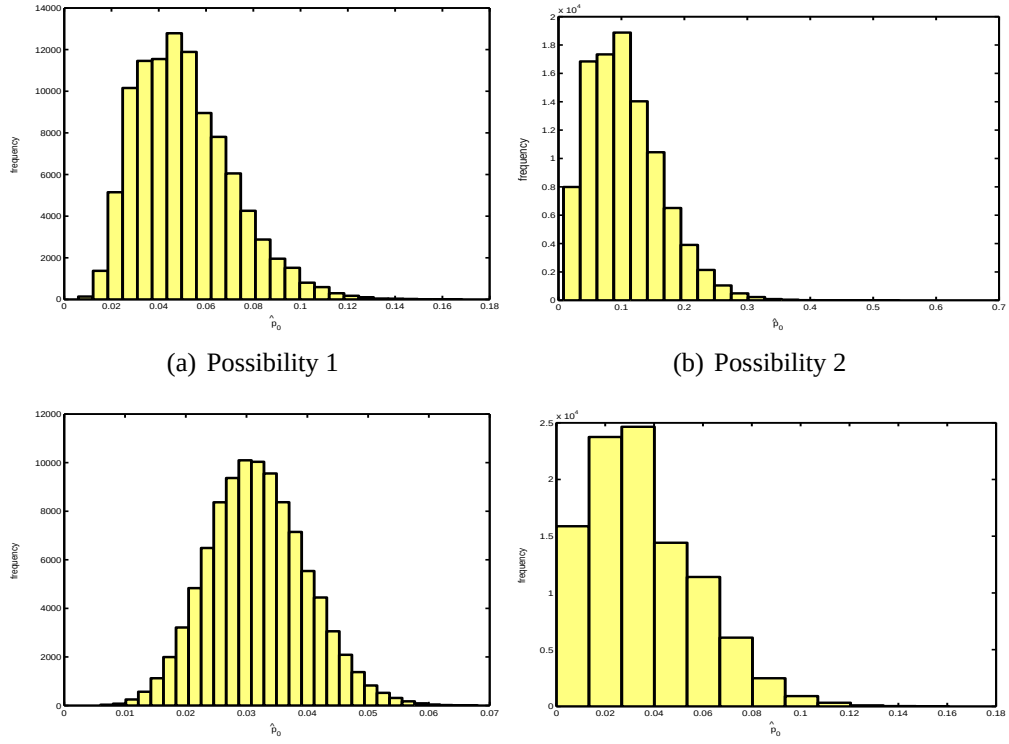
	$\widehat{p}_0$	$\widehat{p}_{0\|0}$	$\widehat{p}_{0\|1}$	$\widehat{p}_{0\|00}$	$\widehat{p}_{0\|01}$	$\widehat{p}_{0\|10}$	$\widehat{p}_{0\|11}$
Possibility 1	0.0510	0.6277	0.0000	1.0000	1.0000	-	0.0000
Possibility 2	0.1068	0.2996	0.0000	1.0000	0.2537	-	0.0000
Possibility 3	0.0320	1.0000	0.0000	1.0000	-	-	0.0196
Possibility 4	0.0350	0.4574	0.0166	1.0000	1.0000	1.0000	0.0000

Table 3.5.3: Point estimates

The point estimate (0.051) of Possibility 1 equals the value for $k = 2$. This is logical since in this possibility the expert fully agrees with the second auditor (who was the expert in the example with two rounds).

In audit controls, the accuracy (and distribution) of the estimator $\widehat{P}_0$ are usually at least as important as the point estimates. Here, they have to be determined by means of simulation, since there are no analytical expressions available. The parameters are assumed to have the estimated values of Table 3.5.3. In the simulation (of 100,000 runs), stratified sampling is applied in such a way that $n_{i_1}^{(2)} > 0$ if $c_{i_1} > 0$. Just as in the example, all possible previously classified 'incorrect' el-

ements are included in the third round. The simulation results in the distributions, presented in Figure 3.5.2.



(a) Possibility 1 (b) Possibility 2

(c) Possibility 3 (d) Possibility 4

Figure 3.5.2: Histograms of simulated distributions of $\widehat{P}_0$

Possibility 3 leads to a fairly symmetrical distribution; the other ones are skewed to the right. The (simulated) standard deviations of $\widehat{P}_0$ are presented in the first line of Table 3.5.4.

	Possibility 1	Possibility 2	Possibility 3	Possibility 4
Three rounds	0.0203	0.0574	0.0079	0.0233
Omission auditor 1	0.0302	0.0591	0.0306	0.0310
Omission auditor 2	0.0320	0.0598	0.0079	0.0350
Omission auditor 1 and 2	0.0490	0.0691	0.0394	0.0409

Table 3.5.4: Standard deviations of $\widehat{P}_0$

The standard deviation is the smallest for Possibility 3, since this is the only case in which no mistakes are found among the classifications of the first auditor.

It is also interesting to look at the accuracy of the estimators, with respect to the design of the repeated audit control. What is the impact of omitting one or several auditors on the accuracy of the estimators? If the first (second) auditor is omitted, the estimator is based on the 53 (500) observations of the second (first) auditor and the 20 observations of the expert. If both internal auditors are omitted, only the 20 flawless observations of the expert are included. Since stratified sampling was used, the estimator $\widehat{P}_0$ is unbiased for all designs. Hence the simulated expectations equal the values $\widehat{p}_0$ in Table 3.5.3. The standard deviations are shown in the last three lines of Table 3.5.4.

Including the observations of all the auditors leads to the smallest standard deviation, while including only the flawless observations of the expert gives the largest standard deviation. Including only one fallible auditor, gives a standard deviation which lies between the previous ones. Omission of either the first or second auditor leads to approximately the same standard deviation for all possibilities, except the third one. In this case, the expert fully agreed with the first auditor: the second did not contribute at all. In the remaining cases, omitting the first auditor leads to a somewhat higher accuracy.

For the Bayesian approach, priors are formulated for P such as described in Section 3.4.3. For our example, the general priors (3.4.4) reduce to the beta distributions

$$\begin{cases} \mathcal{L}(P_1) & = & Beta(\alpha_1, \alpha_0) & \\ \mathcal{L}(P_{1|i_0}) & = & Beta(\alpha_{1|i_0}, \alpha_{0|i_0}), & i_0 = 0, 1 \\ \mathcal{L}(P_{1|i_0 i_1}) & = & Beta(\alpha_{1|i_0 i_1}, \alpha_{0|i_0 i_1}), & i_0, i_1 = 0, 1. \end{cases}$$

To determine the marginal posterior distribution of P_0, the data augmentation procedure is used. For $r = 2$ and $k = 3$, the implementation step (3.4.6) consists of drawing from binomial distributions:

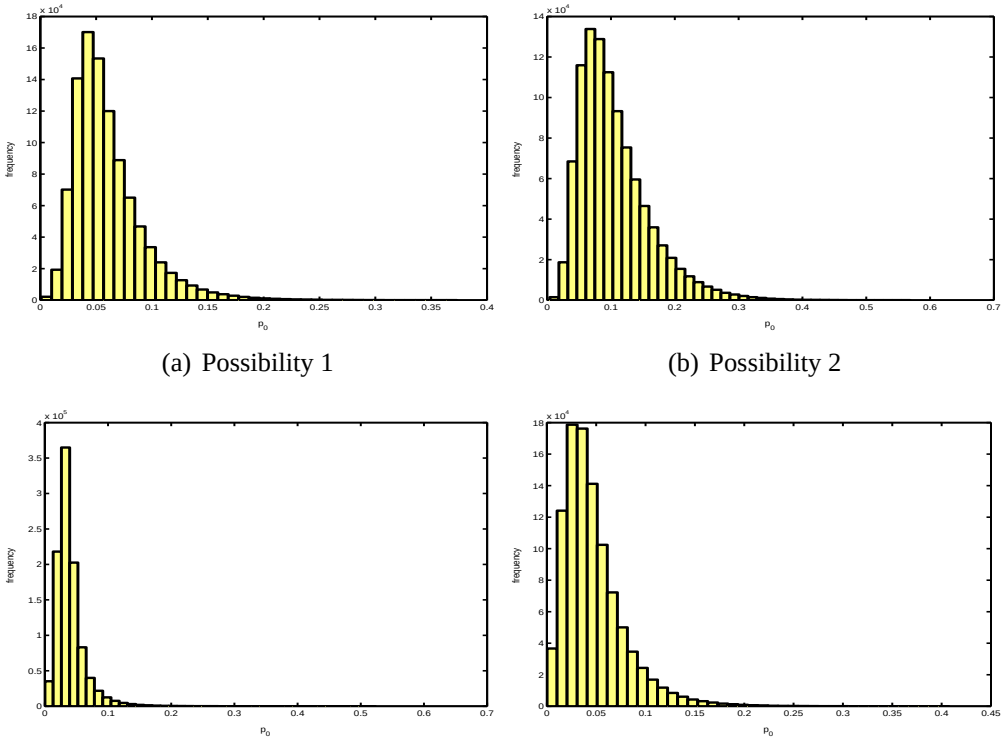
$$\begin{cases} \mathcal{L}(C_{i_1 1}^{(t+1)}) = B(c_{i_1 -}; \pi_{1|i_1}^{(t)}), & i_1 = 0, 1 \\ \mathcal{L}(C_{i_1 i_2 1}^{(t+1)}) = B(c_{i_1 i_2 -} + c_{i_1 i_2}^{(t+1)}; \pi_{1|i_1 i_2}^{(t)}), & i_1, i_2 = 0, 1. \end{cases}$$

The posterior step (3.4.7) reduces to drawing from beta distributions

$$\begin{cases} \mathcal{L}(P_1^{(t+1)}|(\text{simulated})\text{data}) = Beta(\alpha_1 + c_{++1} + c_{++1}^{(t+1)}, \alpha_0 + c_{++0} + c_{++0}^{(t+1)}) \\ \mathcal{L}(P_{1|i_0}^{(t+1)}|(\text{simulated) data}) = \\ \qquad Beta(\alpha_{1|i_0} + c_{1+i_0} + c_{1+i_0}^{(t+1)}, \alpha_{0|i_0} + c_{0+i_0} + c_{0+i_0}^{(t+1)}), \qquad i_0 = 0, 1 \\ \mathcal{L}(P_{1|i_0 i_1}^{(t+1)}|(\text{simulated) data}) = \\ \qquad Beta(\alpha_{1|i_0 i_1} + c_{i_1 1 i_0} + c_{i_1 1 i_0}^{(t+1)}, \alpha_{0|i_0 i_1} + c_{i_1 0 i_0} + c_{i_1 0 i_0}^{(t+1)}), \quad i_0, i_1 = 0, 1. \end{cases}$$

The speed of convergence of the described procedure is related to the fraction of missing observations; since this fraction is very high in our example which has a high dimensionality, the rate of convergence is rather low.

For Jeffrey's noninformative prior (all the α's are 0.5), Figure 3.5.3 shows the marginal posterior distributions for our example, obtained by data augmentation with 1,000,000 iterations:



(a) Possibility 1 (b) Possibility 2

(c) Possibility 3 (d) Possibility 4

Figure 3.5.3: Histograms of simulated posterior distributions of P_0

The mode and 0.95-quantile of the posterior distribution are taken as point estimate and 95%-upper limit. They are presented in the first part of Table 3.5.5.

Jeffrey's noninformative prior				
	Possibility 1	Possibility 2	Possibility 3	Possibility 4
mode	0.043	0.070	0.031	0.029
0.95-quantile	0.121	0.212	0.082	0.107

$\alpha_{0\|0} = \alpha_{1\|1} = 1.5,\ \alpha_{0\|01} = \alpha_{1\|10} = 2.5,\ \alpha_{0\|00} = \alpha_{1\|11} = 3.5$, *other* α *are 0.5*				
	Possibility 1	Possibility 2	Possibility 3	Possibility 4
mode	0.039	0.055	0.030	0.027
0.95-quantile	0.089	0.137	0.057	0.075

Table 3.5.5: Bayesian point estimates and upper limits for p_0

For Jeffrey's prior, the Bayesian point estimates are all smaller than the corresponding classical point estimates (see first column Table 3.5.3).

The second part of Table 3.5.5 contains the estimates for a different prior. The prior parameters are chosen in such a way that the error probabilities of the second fallible auditor are likely to be smaller than those of the first fallible auditor. Moreover, it is more likely that the second auditor's misclassification probabilities are higher if the first auditor has erred previously than if the first auditor gave the correct classification. The impact of this different prior is considerable: especially the upper limits are a lot smaller than for Jeffrey's noninformative prior.

3.6 Conclusions

A general framework for repeated audit controls was introduced for categorical data with $r \geq 2$ levels. Monotone sampling (*cf.* Little and Rubin (2002)) is applied, implying that non-increasing numbers of records are checked by $k \geq 2$ subsequent auditors; the last of these is assumed to be infallible. Two sampling methods were discussed, called random and stratified sampling. In stratified sampling, previous classification results determine the next sample sizes for all classifications separately, while in random sampling they only determine the total

sample size for the next auditor.

It was shown that both sampling methods lead to essentially the same MLE's for the r population fractions p_{i_0}. However, if unstructural zeros occur, the MLE's are not uniquely defined. Since unstructural zeros are much more likely to occur in case of random sampling, we advise stratified sampling for practical use. A further advantage is that the MLE's in this case are unbiased.

A new solution to the unstructural zeros problem was proposed having two advantages: it leads to a MLE with a smaller bias, and encompasses the solutions for the reduced models, where only one error type can occur.

Three different methods to determine upper limits for the fraction incorrect elements in the population were discussed. Of these, the Bayesian approach appeared to be the most satisfactory.

In case error sizes, or relative error sizes (taintings) are observed instead of just error rates, continuous data are obtained. The special case of normally distributed observations with k subsequent auditors is analysed in more detail in the next two chapters. Note that a distribution-free solution can be derived from the present chapter by discretization of the continuous variable into r categories.

3.7 Appendices

3.7.1 Limit distribution

Write

$$\Sigma^{(1)} = Cov(C^{(1)})/n_1$$
$$\Sigma_a^{(j|j-1)} = Cov(C_a^{(j|j-1)})/n_a^{(j)},$$

with elements $f, g = 1, \ldots, r$:

$$\Sigma^{(1)}(f,g) = \begin{cases} \pi_{r-f}(1-\pi_{r-f}) & \text{if } f = g \\ -\pi_{r-f}\pi_{r-g} & \text{if } f \neq g \end{cases}$$

and

$$\Sigma_a^{(j|j-1)}(f,g) = \begin{cases} \pi_{r-f|a}(1-\pi_{r-f|a}) & \text{if } f = g \\ -\pi_{r-f|a}\pi_{r-g|a} & \text{if } f \neq g. \end{cases}$$

Then the asymptotic distributions of the MLE's (3.3.3) are straightforward:

$$\begin{aligned} \sqrt{n_1}vec(\widehat{\Pi}^{(1)} - \Pi^{(1)}) &\xrightarrow{\mathcal{L}} N(0, \Sigma^{(1)}), \\ \sqrt{n_a^{(j)}}vec(\widehat{\Pi}_a^{(j|j-1)} - \Pi_a^{(j|j-1)}) &\xrightarrow{\mathcal{L}} N(0, \Sigma_a^{(j|j-1)}). \end{aligned}$$

If $N_a^{(j)}/n_1 \xrightarrow{P} b_a^{(j)}$ with $b_a^{(j)}$ a constant depending on a,

$$\sqrt{n_1}vec(\widehat{\Pi}_a^{(j|j-1)} - \Pi_a^{(j|j-1)}) \xrightarrow{\mathcal{L}} N(0, \Sigma_a^{(j|j-1)}/b_a^{(j)}).$$

Since $\widehat{\Pi}^{(1)}$ and $\widehat{\Pi}_a^{(j|j-1)}$ are independent, they have an asymptotic multivariate normal distribution with a block-diagonal covariancematrix. The MLE for $\pi_{i_1 \ldots i_k}$ is a function of the preceding estimators (see (3.2.1b)): $\widehat{\Pi}_{i_1 \ldots i_k} = \widehat{\Pi}_{i_1} \cdot \widehat{\Pi}_{i_2|i_1} \cdot \ldots \cdot \widehat{\Pi}_{i_k|i_1 \ldots i_{k-1}}$. Application of the deltamethod (see Lehmann and Casella (1998)) results in the asymptotic distribution of $\widehat{\Pi}_{i_1 \ldots i_k}$. Relation (3.2.1(c)) and applying the deltamethod once more result in the asymptotic distribution of $\widehat{P}_{i_0}$ in (3.3.4).

We only illustrate the whole procedure for the special case $r = k = 2$:

$$\sqrt{n_1}\begin{bmatrix} \widehat{\Pi}_1 - \pi_1 \\ \widehat{\Pi}_0 - \pi_0 \\ \widehat{\Pi}_{1|1} - \pi_{1|1} \\ \widehat{\Pi}_{0|1} - \pi_{0|1} \\ \widehat{\Pi}_{1|0} - \pi_{1|0} \\ \widehat{\Pi}_{0|0} - \pi_{0|0} \end{bmatrix} \xrightarrow{\mathcal{L}} N(0, \Sigma)$$

with

$$\Sigma = \begin{bmatrix} \pi_1\pi_0 & -\pi_1\pi_0 & 0 & 0 & 0 & 0 \\ -\pi_1\pi_0 & \pi_1\pi_0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{\pi_{1|1}\pi_{0|1}}{b_1^{(2)}} & \frac{-\pi_{1|1}\pi_{0|1}}{b_1^{(2)}} & 0 & 0 \\ 0 & 0 & \frac{-\pi_{1|1}\pi_{0|1}}{b_1^{(2)}} & \frac{\pi_{1|1}\pi_{0|1}}{b_1^{(2)}} & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{\pi_{1|0}\pi_{0|0}}{b_0^{(2)}} & \frac{-\pi_{1|0}\pi_{0|0}}{b_0^{(2)}} \\ 0 & 0 & 0 & 0 & \frac{-\pi_{1|0}\pi_{0|0}}{b_0^{(2)}} & \frac{\pi_{1|0}\pi_{0|0}}{b_0^{(2)}} \end{bmatrix}.$$

The deltamethod applied to relation (3.2.1(b)) results in

$$\sqrt{n_1}\begin{bmatrix} \widehat{\Pi}_{11} - \pi_{11} \\ \widehat{\Pi}_{10} - \pi_{10} \\ \widehat{\Pi}_{01} - \pi_{01} \\ \widehat{\Pi}_{00} - \pi_{00} \end{bmatrix} \xrightarrow{\mathcal{L}} N(0, \Delta), \text{ with } \Delta = B'\Sigma B,$$

where $B = \begin{bmatrix} \pi_{1|1} & \pi_{0|1} & 0 & 0 \\ 0 & 0 & \pi_{1|0} & \pi_{0|0} \\ \pi_1 & 0 & 0 & 0 \\ 0 & \pi_1 & 0 & 0 \\ 0 & 0 & \pi_0 & 0 \\ 0 & 0 & 0 & \pi_0 \end{bmatrix}$. So,

$$
\Delta = \begin{bmatrix}
\pi_{1|1}\pi_{11}\pi_0 + \dfrac{\pi_{11}\pi_{10}}{b_1^{(2)}} & \pi_{11}\pi_{0|1}\pi_0 - \dfrac{\pi_{11}\pi_{10}}{b_1^{(2)}} & -\pi_{11}\pi_{01} & -\pi_{11}\pi_{00} \\
\pi_{11}\pi_{0|1}\pi_0 - \dfrac{\pi_{11}\pi_{10}}{b_1^{(2)}} & \pi_{0|1}\pi_{10}\pi_0 + \dfrac{\pi_{11}\pi_{10}}{b_1^{(2)}} & -\pi_{10}\pi_{01} & -\pi_{10}\pi_{00} \\
-\pi_{11}\pi_{01} & -\pi_{10}\pi_{01} & \pi_{1|0}\pi_{01}\pi_1 + \dfrac{\pi_{01}\pi_{00}}{b_0^{(2)}} & \pi_{0|0}\pi_{01}\pi_1 - \dfrac{\pi_{01}\pi_{00}}{b_0^{(2)}} \\
-\pi_{11}\pi_{00} & -\pi_{10}\pi_{00} & \pi_{0|0}\pi_{01}\pi_1 - \dfrac{\pi_{01}\pi_{00}}{b_0^{(2)}} & \pi_{0|0}\pi_{00}\pi_1 + \dfrac{\pi_{01}\pi_{00}}{b_0^{(2)}}
\end{bmatrix}
$$

Applying the deltamethod once again but this time to relation (3.2.1(c)) leads to the asymptotic distribution of $\widehat{P}_{i_0}$: $\sqrt{n_1}\begin{bmatrix} \widehat{P}_1\text{-}p_1 \\ \widehat{P}_0\text{-}p_0 \end{bmatrix} \xrightarrow{\mathcal{L}} N(0, B'\Delta B)$ where

$$B = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

3.7.2 Bias

Derivation of the bias of the modified estimator (3.5.2):

$$\begin{aligned}
E\{\widehat{P}_0^*\} &= E\{E\{\widehat{P}_0^*|N_0^{(2)}\}\} \\
&= Pr(N_0^{(2)}=0)E\{\frac{C_0}{n_1}+\frac{C_1}{n_1}\frac{C_{10}}{N_1^{(2)}}|N_0^{(2)}=0\} \\
&+ \sum_{n_0^{(2)}=1}^{n_2-1} Pr(N_0^{(2)}=n_0^{(2)})E\{\frac{C_0}{n_1}\frac{C_{00}}{N_0^{(2)}}+\frac{C_1}{n_1}\frac{C_{10}}{N_1^{(2)}}|N_0^{(2)}=n_0^{(2)}\} \\
&+ Pr(N_0^{(2)}=n_2)E\{\frac{C_0}{n_1}\frac{C_{00}}{N_0^{(2)}}|N_0^{(2)}=n_2\} \\
&= Pr(N_0^{(2)}=0)(\frac{(n_1-n_2)\pi_0}{n_1}+\frac{(n_1-n_2)\pi_1+n_2}{n_1}\pi_{0|1}) \\
&+ \sum_{n_0^{(2)}=1}^{n_2-1} Pr(N_0^{(2)}=n_0^{(2)})(\frac{(n_1-n_2)\pi_0+n_0^{(2)}}{n_1}\pi_{0|0} \\
&+ \frac{(n_1-n_2)\pi_1+n_1^{(2)}}{n_1}\pi_{0|1}) + Pr(N_0^{(2)}=n_2)(\frac{(n_1-n_2)\pi_0+n_2}{n_1}\pi_{0|0}) \\
&= Pr(N_0^{(2)}=0)(1-\frac{n_2}{n_1})\pi_{01} - Pr(N_0^{(2)}=n_2)(1-\frac{n_2}{n_1})\pi_{10} \\
&+ E\{\frac{(n_1-n_2)\pi_0+N_0^{(2)}}{n_1}\pi_{0|0}+\frac{(n_1-n_2)\pi_1+N_1^{(2)}}{n_1}\pi_{0|1}\} \\
&= p_0+(1-\frac{n_2}{n_1})(\pi_1^{n_2}\pi_{01}-\pi_0^{n_2}\pi_{10}).
\end{aligned}$$

The bias of $\widehat{P}_0$ (3.5.1) can be derived in a similar way.

Chapter 4

Multivariate regression

4.1 Introduction

In this chapter - and the next - the perspective broadens: instead of categorical variables, continuous variables will be considered. Besides, we temporarily leave the specific auditing problem and direct our attention to a very general situation: we consider multivariate regression where new dependent variables are consecutively added during the experiment (or in time). Since, no retrospective observations are assumed to be possible, the number of observations decreases with the added variables. The explanatory variables are observed throughout.

Two examples will illustrate this set-up. The first considers male patients who receive a new cholesterol decreasing medicine. The explanatory variables are age, weight and medication. First, only the decrease in cholesterol is observed; for later patients, pulse and blood pressure as well, and still later haemoglobine is measured. The second example relates to a chemical process, where the quantities of three main ingredients are used as the explanatory variables. In the beginning, the only variable observed on consecutive days is the quantity of produced material. Later the production of two by-products is measured as well, and finally also the CO_2 emission.

In Section 4.2 the model is presented in detail and illustrated with a numerical example. In Section 4.3, four classical estimation procedures are discussed: O(rdinary) L(east) S(quares), G(eneralized) LS, E(stimated) GLS and ML. For

LS estimation, only assumptions about the first two moments are required; for ML estimation, we assume normality. As to all regression coefficients, it is shown that a specific choice of EGLS coincides with ML. All estimators appear to have a clear geometric interpretation. Section 4.4 discusses the relative efficiency of the OLS estimators in relation to the (E)GLS estimators.

The model with complete observations follows as a special case. The same holds for the model with the constant term as one of the explanatory variables, leading to centered variables. Both cases are treated in Section 4.5. Section 4.6 describes estimation under linear restrictions and gives MANOVA-tables to perform exact L(ikelihood) R(atio) tests on the coefficients. Section 4.7 reviews the Wishart and Wilks' distribution and introduces a generalized Wilks' distribution that gives the exact distribution of our test statistics in Section 4.8. In Section 4.9 the presented estimation and testing techniques are applied to a numerical example. In Section 4.10 several χ^2-approximations of the generalized Wilks' distribution are derived and compared by means of simulation. The final Section 4.11 contains the main conclusions and ideas for future research.

The perspective of the problem can be reversed: instead of regarding the observations of the newly added variables as additional information, the lacking past observations of these new variables can be regarded as missing data. Practical examples of this type of monotone missing data patterns are panel surveys with either drop outs or new members. However, the linear regression model and its analysis only hold under very strict conditions for the missing data mechanism; an example of this is missing completely at random, see Rubin (1976).

To solve missing data problems, general techniques are multiple imputation, data augmentation and the E(xpectation)M(aximization)-algorithm. The EM-algorithm is a widely used technique to determine ML estimates in missing data problems. Although this algorithm converges to ML estimates, it does not give analytical closed-form expressions for the estimators, nor does it lead to exact distributions of test statistics. Therefore, our approach is much simpler and more straightforward.

The model with only the constant term as explanatory variable has received a lot of attention in the missing data literature; see Little and Rubin (2002) for an

overview. Under the assumption of normality, observations missing at random, and distinctness (see Rubin (1976)), several authors derived the MLE's by means of factorization of the likelihood or tedious matrix differentiation. Our formulae contain these previous results as a very special case, see Section 4.5.2.

Finally, we mention that our general case of multivariate regression with missing observations of the dependent variables was considered in Robins and Rotnitzky (1995), who discuss semiparametric asymptotic efficiency.

4.2 The model

Consider the multivariate linear regression model with M dependent variables and k (deterministic) explanatory variables; observations are gathered for N cases. Let $X_{tj} \in I\!R$ be the observed value of the j^{th} explanatory variable ($j = 1, \ldots, k$) for the t^{th} case; complete data are available for the explanatory variables, so $t = 1, \ldots, N$ for all j.

The observations of the dependent variables are incomplete; the dependent variables are ordered such that later added variables come last. So their data are divided into r ordered groups according to the pattern of increasingly missing data. Group i contains m_i variables for which exactly the first N_i observations are available:

$$N = N_1 \geq N_2 \geq \ldots \geq N_r; \quad M_i = \sum_{j=1}^{i} m_j \quad (i = 1, \ldots, r,\ M_r = M).$$

The vector $Y_{ti} \in I\!R^{m_i}$ contains the values of these m_i dependent variables for case t. So Y_{ti} is observable for $t = 1, \ldots, N_i$ and missing for $t = N_i + 1, \ldots, N$. The special case $N = N_1 = \ldots = N_r$ gives the usual complete model.

The r (multivariate) regression equations can be written as

$$Y_{ti} = \mu_{ti} + \varepsilon_{ti}, \quad \mu_{ti} = \sum_{j=1}^{k} X_{tj}\beta_{ji}, \quad i = 1, \ldots, r, \quad t = 1, \ldots, N_i, \quad (4.2.1)$$

where $\beta_{ji} \in I\!R^{m_i}$ denotes a vector of unknown regression coefficients. For the errors we assume

$$E\{\varepsilon_{ti}\} = 0, \quad Cov(\varepsilon_{ti}, \varepsilon_{sj}) = \delta_{ts}\sigma_{ij}, \quad (4.2.2)$$

with (completely unknown) non-singular $\Sigma = (\sigma_{ij}) \in I\!R^{M\times M}$ not depending on the β_{ji}. We write $\Sigma > 0$ for positive definiteness. If normality of the errors is assumed, it will be mentioned explicitly.

The union of the groups 1 up to i will be denoted by (i), hence $Y_{t(i)} = (Y'_{t1} \dots Y'_{ti})' \in I\!R^{M_i}$, $i = 1, \dots, r$ and similarly for $\mu_{t(i)}$ and $\varepsilon_{t(i)}$.

The OLS criterion is simply minimizing the sum of squares of the errors, which can be written as ($N_{r+1} := 0$) :

$$\sum_{i=1}^{r} \sum_{t=N_{i+1}+1}^{N_i} \varepsilon'_{t(i)} \varepsilon_{t(i)}. \tag{4.2.3}$$

The solution of this minimization problem w.r.t. the β_{ji} will be given in Section 4.3.2.

The GLS criterion is minimizing the weighted sum of squares with the inverse of the covariance matrix of all errors as weight matrix. Since errors of different cases are uncorrelated, it can be written in a more simple form. The error covariance matrix $\Sigma_{(i)(i)}$ of $\varepsilon_{t(i)}$ can be partitioned as follows

$$\Sigma_{(i)(i)} := Cov(\varepsilon_{t(i)}) = Cov \begin{pmatrix} \varepsilon_{t(i-1)} \\ \varepsilon_{ti} \end{pmatrix} = \begin{bmatrix} \Sigma_{(i-1)(i-1)} & \Sigma_{(i-1)i} \\ \Sigma_{i(i-1)} & \Sigma_{ii} \end{bmatrix}. \tag{4.2.4}$$

So, $\Sigma_{(i)(i)} \in I\!R^{M_i \times M_i}$, $\Sigma_{(i-1)(i-1)} \in I\!R^{M_{i-1} \times M_{i-1}}$, $\Sigma_{(i-1)i} \in I\!R^{M_{i-1} \times m_i}$ and in particular $\Sigma_{(r)(r)} = \Sigma$ and $\Sigma_{(1)(1)} = \Sigma_{11}$. Then, using (4.2.4), the GLS criterion can be written as

$$\sum_{i=1}^{r} \sum_{t=N_{i+1}+1}^{N_i} \varepsilon'_{t(i)} \Sigma^{-1}_{(i)(i)} \varepsilon_{t(i)}. \tag{4.2.5}$$

This minimization problem w.r.t. the β_{ji} will be treated in Section 4.3.3. In contrast with the complete model GLS and OLS no longer coincide. Since GLS is BLUE, it outperforms OLS.

Of course, in practice Σ is unknown and GLS cannot be applied. In Section 4.3.4 we therefore consider EGLS estimation, where Σ is replaced by some estimator. We discuss shortly several possible estimators. One specific choice is analysed in detail. In Section 4.3.5 we consider ML estimation under normality;

it will be shown that the specific form of EGLS estimation coincides with ML estimation.

Numerical illustration

The notations are illustrated by means of the following fictitious data, related to the examples of Section 4.1 with four dependent and three explanatory variables (excluding the constant). As usual, columns of X (and Y) refer to variables and rows to cases. Not observed values in Y are denoted by parentheses. We nevertheless give these values to compare the results obtained from the incomplete data with the results for the complete data.

$$X = \begin{bmatrix} 1 & 5 & 5 & 7 \\ 1 & 1 & 3 & 1 \\ 1 & 3 & 3 & 1 \\ 1 & 3 & 1 & 3 \\ 1 & 5 & 5 & 7 \\ 1 & 1 & 3 & 1 \\ 1 & 3 & 3 & 1 \\ 1 & 3 & 1 & 3 \\ 1 & 4 & 4 & 5 \\ 1 & 2 & 3 & 2 \\ 1 & 3 & 3 & 2 \\ 1 & 3 & 2 & 3 \end{bmatrix} \qquad Y = \begin{bmatrix} 7 & 5 & 6 & 1 \\ 5 & 9 & 2 & 4 \\ 7 & 5 & 10 & 6 \\ 1 & 1 & 2 & 5 \\ 4 & 2 & 0 & 4 \\ 5 & 9 & 8 & 4 \\ 7 & 8 & 4 & 6 \\ 4 & 1 & 8 & 2 \\ 3 & 2 & 4 & 1 \\ 5 & 7 & 5 & 4 \\ 6 & 8 & 6 & (5) \\ 6 & (3) & (5) & (6) \end{bmatrix}$$

$$\begin{array}{l} r = 3 \\ k = 4 \\ \\ N_1 = 12 \\ N_2 = 11 \\ N_3 = 10 \\ \\ m_1 = 1 \\ m_2 = 2 \\ m_3 = 1 \\ \\ M_1 = 1 \\ M_2 = 3 \\ M_3 = 4 \end{array}$$

So, for example,

$$X_{1,4} = 7, \quad Y_{1,1} = 7, \quad Y_{1,2} = \begin{bmatrix} 5 \\ 6 \end{bmatrix}, \quad Y_{1(2)} = \begin{bmatrix} 7 \\ 5 \\ 6 \end{bmatrix}, \quad Y_{1,3} = 1,$$

and (4.2.1) reads for $i = 2$:

$$Y_{t,2} = \beta_{1,2} + X_{t,2}\beta_{2,2} + X_{t,3}\beta_{3,2} + X_{t,4}\beta_{4,2} + \varepsilon_{t,2}, \quad t = 1, \ldots, 11.$$

Note that suffices are separated by a comma whenever confusion threatens.

4.3 Estimation

4.3.1 Notation

We introduce some column- and matrix-notation for the observed variables and regression coefficients. The index i refers to group i and (i) again to the union of the groups $1, 2, \ldots, i$.

$$X = \begin{bmatrix} \boxed{\begin{matrix} X_{1,1} & X_{1,2} & \cdots & X_{1,k} \\ \vdots & \vdots & & \vdots \\ X_{N_i,1} & X_{N_i,2} & \cdots & X_{N_i,k} \end{matrix}} \\ \vdots \\ X_{N,1} \quad X_{N,2} \quad \cdots \quad X_{N,k} \end{bmatrix}$$

$$\begin{bmatrix} \boxed{X_i} \end{bmatrix}$$

$$\beta = \begin{bmatrix} \boxed{\begin{matrix} \beta'_{1,1} & \cdots & \beta'_{1,i-1} \\ \vdots & & \vdots \\ \beta'_{j,1} & \cdots & \beta'_{j,i-1} \\ \vdots & & \vdots \\ \beta'_{k,1} & \cdots & \beta'_{k,i-1} \end{matrix}} & \begin{matrix} \beta'_{1,i} & \cdots & \beta'_{1,r} \\ \vdots & & \vdots \\ \beta'_{j,i} & \cdots & \beta'_{j,r} \\ \vdots & & \vdots \\ \beta'_{k,i} & \cdots & \beta'_{k,r} \end{matrix} \end{bmatrix}$$

$$\begin{bmatrix} \boxed{\beta_{(i-1)}} & \beta_i & \cdots & \beta_r \end{bmatrix}$$

So $X_i \in I\!R^{N_i \times k}$ is the matrix with the first N_i observations of all explanatory variables. The submatrices $\beta_{(i-1)} \in I\!R^{k \times M_{i-1}}$ and $\beta_i \in I\!R^{k \times m_i}$ of $\beta \in I\!R^{k \times M}$ contain the regression coefficients corresponding to groups $(i-1)$ and i of dependent variables, respectively. The Y_{ti} can be grouped in a corresponding way:

$$\begin{bmatrix} \boxed{\begin{matrix} Y'_{1,1} & \cdots & Y'_{1,i-1} \\ \vdots & & \vdots \\ \vdots & & \vdots \\ Y'_{N_i,1} & \cdots & Y'_{N_i,i-1} \end{matrix}} & \begin{matrix} Y'_{1,i} & \cdots & Y'_{1,r} \\ \vdots & \cdots & \vdots \\ \vdots & \cdots & Y'_{N_r,r} \\ Y'_{N_i,i} & & \end{matrix} \\ \begin{matrix} \vdots & \cdots & Y'_{N_{i-1},i-1} \\ Y'_{N,1} & & \end{matrix} & \end{bmatrix}$$

$$\begin{bmatrix} \boxed{Y_{(i-1)}} & Y_i & \cdots & Y_r \end{bmatrix}$$

The matrix $Y_i \in I\!R^{N_i \times m_i}$ contains all observations of group i. But the matrix $Y_{(i-1)} \in I\!R^{N_i \times M_{i-1}}$ contains *only* the first N_i observations of the foregoing groups $(i-1)$ (with $Y_{(0)} = 0$). We use similar definitions for the μ_{ti} and ε_{ti}.

4.3.2 OLS estimation

From (4.2.1) we get $(i = 1, \ldots, r)$

$$\begin{cases} Y_i = \mu_i + \varepsilon_i, & \mu_i = X_i\beta_i \\ Y_{(i-1)} = \mu_{(i-1)} + \varepsilon_{(i-1)}, & \mu_{(i-1)} = X_i\beta_{(i-1)}. \end{cases} \tag{4.3.1}$$

Then the OLS criterion (4.2.3) can be written as

$$\sum_{i=1}^{r} \text{tr}(\varepsilon_i'\varepsilon_i). \tag{4.3.2}$$

So the OLS estimates can be found by columnwise orthogonal projections. We define the following relevant spaces and accompanying characteristics:

$$\begin{array}{lcll} L_i & = & \mathcal{R}(X_i) & \text{: the space spanned by the columns of } X_i, \\ H_i & \in & I\!R^{N_i \times N_i} & \text{: the orthogonal projection matrix of } L_i, \\ U_i & = & I_{N_i} - H_i & \text{: the orthogonal projection matrix of } L_i^{\perp}, \\ l_i & = & dim(L_i) = r(X_i), \quad r_i = dim(L_i^{\perp}) = N_i - l_i. & \end{array}$$

Clearly each column of μ_i is element of L_i. To indicate this property, we will use the (short) notation $\mu_i \in L_i$.

Theorem 4.3.1. *The OLS estimator for μ_i $(i = 1, \ldots, r)$ is the (columnwise) orthogonal projection of Y_i onto $\mathcal{R}(X_i)$:*

$$Z_i := H_i Y_i. \tag{4.3.3}$$

Proof. The OLS criterion (4.3.2) is the sum of r squared lengths of the error terms. Since the mean μ_i only appears in the i^{th} term, (4.3.2) is minimized by minimization of these terms separately. With respect to term i we can write

$$\varepsilon_i = Y_i - \mu_i = H_i(Y_i - \mu_i) + U_i(Y_i - \mu_i) = (Z_i - \mu_i) + U_i Y_i.$$

Clearly, the minimum is attained for $\mu_i = Z_i$. □

The OLS estimator for ε_i ($i = 1, \ldots, r$) follows from relations (4.3.1) and (4.3.3):

$$E_i = Y_i - Z_i = U_i Y_i = U_i \varepsilon_i. \tag{4.3.4}$$

OLS estimators b_i for the regression coefficients β_i are given by

$$b_i = G_i X_i' Y_i \quad \text{with } G_i = (X_i' X_i)^-, \tag{4.3.5}$$

where a g-inverse is denoted by $^-$. It is clear that the OLS estimators b_i are unbiased in case of non-collinearity.

We use the notation $E_{(i-1)g}$ for the columns $M_{g-1} + 1$ through M_g of $E_{(i-1)}$, *i.e.* the first N_i rows of OLS residuals corresponding to group g (a similar notation is used for the error-terms, (E)GLS residuals, *et cetera*). Similarly to (4.3.4), we have

$$U_i E_{(i-1)g} = U_i (Y_{(i-1)g} - Z_{(i-1)g}) = U_i Y_{(i-1)g} = U_i \varepsilon_{(i-1)g}. \tag{4.3.6}$$

We propose the following estimator for the covariance matrix Σ

$$S_{ii} = E_i' E_i / r_i, \quad S_{ig} = E_i' E_{(i-1)g} / r_i \quad \text{for } g = 1, \ldots, i-1. \tag{4.3.7}$$

This estimator S is unbiased for Σ because S_{ii} and S_{ig} are unbiased for σ_{ii} and σ_{ig}, respectively. Without loss of generality we take $m_i = 1$ for all i, so the unbiasedness of S_{ig} follows from

$$E\{E_i' E_{(i-1)g}\}/r_i = tr\left(U_i E\{\varepsilon_{(i-1)g}' \varepsilon_i\}\right)/r_i = \frac{\sigma_{ig}}{r_i} tr(U_i) = \sigma_{ig},$$

where the second equality is based on (4.3.4) and (4.3.6).

Whether S is positive semidefinite depends on the relative difference between the group sample sizes N_i. If the relative difference is small, S will tend to be positive semidefinite. To ensure that a positive semidefinite S is even positive definite, we impose the regularity condition $N_r \geq M_r + l_r$.

4.3.3 GLS estimation

GLS estimation is usually only of theoretical interest, because in practice the covariance matrix Σ is unknown. However, GLS estimators are BLUE and outperform the OLS estimators in this sense. So we may hope to do better than OLS by replacing Σ in the formulae for GLS with a suitable estimator $\widehat{\Sigma}$ (EGLS, see Section 4.3.4).

We rewrite the GLS criterion (4.2.5) in a form more suitable for minimization. Let

$$\begin{cases} \alpha_i := \Sigma^{-1}_{(i-1)(i-1)}\Sigma_{(i-1)i} \in \mathbb{R}^{M_{i-1}\times m_i} \\ \zeta_{ti} := \alpha_i'\varepsilon_{t(i-1)} \in \mathbb{R}^{m_i\times 1} \\ \eta_{ti} := \varepsilon_{ti} - \zeta_{ti} \in \mathbb{R}^{m_i\times 1} \\ \nu_{ti} := \mu_{ti} + \zeta_{ti} \in \mathbb{R}^{m_i\times 1}. \end{cases} \tag{4.3.8}$$

Note that $Y_{t(0)} = \varepsilon_{t(0)} = 0$, so $\zeta_{t1} = 0$, $\eta_{t1} = \varepsilon_{t1}$ and $\nu_{t1} = \mu_{t1}$. Then $\eta_{t1}, \ldots, \eta_{tr}$ are uncorrelated, η_{ti} and ν_{ti} are uncorrelated and

$$\begin{cases} E\{\zeta_{ti}\} = E\{\eta_{ti}\} = 0 \\ \Delta_{ii} := Cov(\zeta_{ti}) = \alpha_i'\Sigma_{(i-1)(i-1)}\alpha_i \\ \Gamma_{ii} := Cov(\eta_{ti}) = \Sigma_{ii} - \Delta_{ii}. \end{cases} \tag{4.3.9}$$

In case of normality we have the interpretation

$$\begin{cases} \nu_{ti} = E\{Y_{ti}|Y_{t(i-1)}\} \\ \Gamma_{ii} = Cov(Y_{ti}|Y_{t(i-1)}). \end{cases} \tag{4.3.10}$$

From (4.3.8), (4.3.9) we get

$$\Sigma^{-1}_{(i)(i)} = \begin{bmatrix} \Sigma^{-1}_{(i-1)(i-1)} + \alpha_i\Gamma^{-1}_{ii}\alpha_i' & -\alpha_i\Gamma^{-1}_{ii} \\ -\Gamma^{-1}_{ii}\alpha_i' & \Gamma^{-1}_{ii} \end{bmatrix}$$

and so

$$\varepsilon'_{t(i)}\Sigma^{-1}_{(i)(i)}\varepsilon_{t(i)} = \varepsilon'_{t(i-1)}\Sigma^{-1}_{(i-1)(i-1)}\varepsilon_{t(i-1)} + \eta'_{ti}\Gamma^{-1}_{ii}\eta_{ti}.$$

Therefore, the GLS criterion (4.2.5) can be rewritten as

$$\sum_{i=1}^{r}\sum_{t=1}^{N_i}\eta'_{ti}\Gamma^{-1}_{ii}\eta_{ti}. \tag{4.3.11}$$

For the ζ_{ti}, η_{ti} and ν_{ti} we use the same block notation as for the Y_{ti} (and μ_{ti} and ε_{ti}; see Section 4.2). So from (4.3.8) $(i = 1, \ldots, r)$

$$\begin{cases} Y_i = \nu_i + \eta_i \\ \nu_i = \mu_i + \zeta_i \\ \zeta_i = \varepsilon_{(i-1)}\alpha_i = Y_{(i-1)}\alpha_i - \mu_{(i-1)}\alpha_i \\ \varepsilon_i = \zeta_i + \eta_i. \end{cases} \tag{4.3.12}$$

The GLS criterion (4.3.11) can be written as

$$\textstyle\sum_{i=1}^{r} \mathrm{tr}(\Gamma_{ii}^{-1}\eta_i'\eta_i). \tag{4.3.13}$$

This form leads to the solution in Theorem 4.3.2.

Theorem 4.3.2. *The GLS estimator for* μ_i $(i = 1, \ldots, r)$ *is*

$$\breve{\mu}_i := H_i(Y_i - Y_{(i-1)}\alpha_i + \breve{\mu}_{(i-1)}\alpha_i), \text{ with } \breve{\mu}_{(0)} := 0. \tag{4.3.14}$$

Proof. The GLS criterion (4.3.13) is a summation over all groups. Clearly the mean μ_i not only appears in the i^{th} term but also in all subsequent terms $i + 1, \ldots, r$. So minimization of (4.3.13) has to take place in an sequential way, starting with group r. Since μ_i, $\mu_{(i-1)} \in L_i$ we get with (4.3.12):

$$\begin{aligned} \eta_i &= Y_i - \nu_i = Y_i - Y_{(i-1)}\alpha_i + \mu_{(i-1)}\alpha_i - \mu_i \\ &= H_i(Y_i - Y_{(i-1)}\alpha_i + \mu_{(i-1)}\alpha_i - \mu_i) + U_i(Y_i - Y_{(i-1)}\alpha_i + \mu_{(i-1)}\alpha_i - \mu_i) \\ &= \big[H_i(Y_i - Y_{(i-1)}\alpha_i + \mu_{(i-1)}\alpha_i) - \mu_i\big] + U_i(Y_i - Y_{(i-1)}\alpha_i). \end{aligned}$$

Regardless of the value of Γ_{rr} and given $\mu_{(r-1)}$, the first term of this orthogonal decomposition of η_r is zero for $\mu_r = H_r(Y_r - Y_{(r-1)}\alpha_r + \mu_{(r-1)}\alpha_r)$. After substituting this minimum into (4.3.13), μ_{r-1} only appears in the $(r-1)^{th}$ term, *et cetera*. Since $Y_{(i-1)} = \mu_{(i-1)} = 0$ for $i = 1$, repeated application of the preceding argumentation results in the closed form GLS estimator (4.3.14). □

Relation (4.3.1) and $\breve{\mu}_i$ given by (4.3.14) lead to the GLS estimator $\breve{\varepsilon}_i$ for ε_i. Next, the GLS estimators $\breve{\zeta}_i$, $\breve{\nu}_i$ and $\breve{\eta}_i$ for ζ_i, ν_i and η_i, respectively, follow from relation (4.3.12).

From expression (4.3.14), it is clear that the GLS estimates have to be determined sequentially, *i.e.* only after the GLS estimates for group $i-1$ are determined, it is possible to determine the estimates for group i. So the GLS estimators in the proof of Theorem 4.3.2 are derived sequentially starting with the last group, while the actual estimates are determined sequentially starting with the first group.

The definitions (4.3.3) and (4.3.14) immediately imply the next Corollary.

Corollary. *The GLS estimators $\breve{\mu}_i$ and $\breve{\varepsilon}_i$ can be written in relation to the OLS estimators Z_i and E_i as*

$$\begin{cases} \breve{\mu}_i = Z_i - H_i \breve{\zeta}_i, \quad \textit{with } \breve{\zeta}_i := \breve{\varepsilon}_{(i-1)}\alpha_i \\ \breve{\varepsilon}_i = E_i + H_i \breve{\zeta}_i. \end{cases} \tag{4.3.15}$$

Since the GLS estimators $\breve{\mu}_i$ are the (columnwise) orthogonal projections of $Y_i - \breve{\zeta}_i$ onto $\mathcal{R}(X_i)$, it follows that $X_i\breve{\beta}_i = \breve{\mu}_i = H_i(Y_i - \breve{\zeta}_i)$. So, GLS estimators $\breve{\beta}_i$ for β_i $(i = 1, \ldots, r)$ are given by

$$\breve{\beta}_i = G_i X_i'(Y_i - \breve{\zeta}_i). \tag{4.3.16}$$

The GLS estimators $\breve{\mu}_i$ are BLUE. So the $\breve{\beta}_i$ are BLUE for estimable β_i.

The achieved minimum of the GLS criterion (4.2.5), (4.3.11) or (4.3.13) is

$$\sum_{i=1}^{r} \text{tr}(\Gamma_{ii}^{-1}\breve{\eta}_i'\breve{\eta}_i). \tag{4.3.17}$$

4.3.4 EGLS estimation

In the more common situation in which both the regression coefficients and the covariance matrix are unknown, EGLS is often applied. For EGLS we have to minimize (4.2.5), where the covariance-matrix Σ is replaced by an estimate, for example the OLS estimator S of (4.3.7). We will consider here another, more implicitly defined estimator for Σ as well. (In Section 4.3.5 we will see the relation with ML.)

Note that estimation of Σ is equivalent to estimation of (α_i, Γ_{ii}) $i = 1, \ldots, r$. From the expressions (4.3.14) for the GLS estimators $\breve{\mu}_i$ it is clear that they depend on the α_i but not on the Γ_{ii}. So only the EGLS estimators $\widehat{\alpha}_i$ for the α_i are

relevant for the EGLS estimators $\widehat{\mu}_i$ for μ_i; they do not depend on the choices $\widehat{\Gamma}_{ii}$ for Γ_{ii}.

Now we take a very specific choice of the $\widehat{\alpha}_i$, leaving the Γ_{ii} undetermined for the moment. We define our $\widehat{\alpha}_i$ as minimizing (4.3.17). Clearly, this is equivalent to minimizing (4.3.13) simultaneously to α_i and β_i. For this minimization problem, we consider orthogonal projections onto extended spaces $L_{(i)} \supseteq L_i$. We define

$$\begin{array}{lcll} L_{(i)} & = & \mathcal{R}(X_i\, Y_{(i-1)}) & = L_i \oplus \mathcal{R}(Y_{(i-1)}), \text{ (with } Y_{(0)} := 0), \\ H_{(i)} & \in & \mathbb{R}^{N_i \times N_i} & : \text{orthogonal projection matrix of } L_{(i)}, \\ U_{(i)} & = & I_{N_i} - H_{(i)} & : \text{orthogonal projection matrix of } L_{(i)}^{\perp}, \\ l_{(i)} & = & dim(L_{(i)}), & r_{(i)} = dim(L_{(i)}^{\perp}) = N_i - l_{(i)}. \end{array}$$

Since $\mathcal{R}(X_i) \cap \mathcal{R}(Y_{(i-1)}) = \{0\}$ a.s., any $\nu_i \in L_{(i)}$ can be uniquely written as $\nu_i = \mu_i + \zeta_i$, with $\mu_i \in \mathcal{R}(X_i)$ and $\zeta_i \in \mathcal{R}(Y_{(i-1)})$. Note that μ_i is the (oblique) projection of ν_i onto $\mathcal{R}(X_i)$ along $\mathcal{R}(Y_{(i-1)})$, and that ζ_i is the (oblique) projection of ν_i onto $\mathcal{R}(Y_{(i-1)})$ along $\mathcal{R}(X_i)$. We call shortly μ_i the $\mathcal{R}(X_i)$-projection of ν_i, and ζ_i the $\mathcal{R}(Y_{(i-1)})$-projection of ν_i.

Theorem 4.3.3. *The EGLS estimator for μ_i is the $\mathcal{R}(X_i)$-projection of $\widehat{\nu}_i$, where $\widehat{\nu}_i$ is the EGLS estimator for ν_i given by*

$$\widehat{\nu}_i := H_{(i)} Y_i. \tag{4.3.18}$$

Proof. The EGLS estimator for ν_i follows straightforwardly from orthogonal decompositions (compare the proof of Theorem 4.3.2). Since $\nu_i \in L_{(i)}$ we have:

$$\eta_i = Y_i - \nu_i = H_{(i)}(Y_i - \nu_i) + U_{(i)}(Y_i - \nu_i) = (H_{(i)} Y_i - \nu_i) + U_{(i)} Y_i.$$

So, the EGLS estimator for ν_i is given by (4.3.18) regardless of Γ_{ii}. Since $\widehat{\nu}_i \in L_{(i)}$, $\widehat{\mu}_i \in \mathcal{R}(X_i)$ and $\widehat{\zeta}_i \in \mathcal{R}(Y_{(i-1)})$, we see that $\widehat{\mu}_i$ is the $\mathcal{R}(X_i)$-projection of $\widehat{\nu}_i$. □

Note that the proof implies that $\widehat{\zeta}_i$ is the $\mathcal{R}(Y_{(i-1)})$-projection of $\widehat{\nu}_i$. Relation (4.3.12) and $\widehat{\nu}_i$ lead to the EGLS estimators $\widehat{\eta}_i$ for η_i, and $\widehat{\varepsilon}_i$ for ε_i.

The property $H_i \widehat{\eta}_i = H_i(Y_i - \widehat{\zeta}_i - \widehat{\mu}_i) = 0$ immediately gives the next Corollary.

Corollary. *The EGLS estimators $\widehat{\mu}_i$ and $\widehat{\varepsilon}_i$ for μ_i and ε_i, respectively, can be written in relation to the OLS estimators Z_i and E_i as*

$$\begin{cases} \widehat{\mu}_i = Z_i - H_i\widehat{\zeta}_i, \\ \widehat{\varepsilon}_i = E_i + H_i\widehat{\zeta}_i. \end{cases} \tag{4.3.19}$$

Since $Y_{(i-1)} = \widehat{\mu}_{(i-1)} + \widehat{\varepsilon}_{(i-1)}$ and $\widehat{\mu}_{(i-1)} \in \mathcal{R}(X_i)$, we have $L_{(i)} := \mathcal{R}(X_i\ Y_{(i-1)}) = \mathcal{R}(X_i\ \widehat{\varepsilon}_{(i-1)})$ and so $\widehat{\zeta}_i$ is the $\mathcal{R}(\widehat{\varepsilon}_{(i-1)})$-projection of $\widehat{\nu}_i = H_{(i)}Y_i$. To obtain simple expressions, we will make use of projections onto $\mathcal{R}(\widehat{\varepsilon}_{(i-1)})$ instead of $\mathcal{R}(Y_{(i-1)})$. Since

$$\widehat{\nu}_i = \widehat{\mu}_i + \widehat{\zeta}_i = X_i\widehat{\beta}_i + \widehat{\varepsilon}_{(i-1)}\widehat{\alpha}_i = \begin{bmatrix} X_i\ \widehat{\varepsilon}_{(i-1)} \end{bmatrix} \begin{bmatrix} \widehat{\beta}_i \\ \widehat{\alpha}_i \end{bmatrix},$$

EGLS estimators $(\widehat{\beta}_i,\ \widehat{\alpha}_i)$ for $(\beta_i,\ \alpha_i)$, are given by

$$\begin{bmatrix} \widehat{\beta}_i \\ \widehat{\alpha}_i \end{bmatrix} = G_{(i)} \begin{bmatrix} X_i' \\ \widehat{\varepsilon}_{(i-1)}' \end{bmatrix} Y_i, \text{ with } G_{(i)} = \begin{bmatrix} X_i'X_i & X_i'\widehat{\varepsilon}_{(i-1)} \\ \widehat{\varepsilon}_{(i-1)}'X_i & \widehat{\varepsilon}_{(i-1)}'\widehat{\varepsilon}_{(i-1)} \end{bmatrix}^{-}. \tag{4.3.20}$$

Since $\widehat{\varepsilon}_{(0)} = 0$, we can always take $\widehat{\beta}_1 = b_1$ given by (4.3.5).

In case of normally distributed errors $E\{Y_i|Y_{(i-1)}\} = \nu_i$, hence $E\{\widehat{\nu}_i|Y_{(i-1)}\} = H_{(i)}E\{Y_i|Y_{(i-1)}\} = H_{(i)}\nu_i = \nu_i$. Since $\widehat{\mu}_i$, (μ_i) is an (oblique) projection of $\widehat{\nu}_i$ (ν_i) onto L_i, it follows that $E\{\widehat{\mu}_i|Y_{(i-1)}\} = \mu_i$ (see Malinvaud (1970) *e.g.*). If $r(X_i) = k$, there is a one-to-one linear relationship between μ_i and β_i, so $\widehat{\beta}_i$ is unbiased as well.

The geometric interpretations and the underlying relations of the OLS and EGLS estimators are shown in Figure 4.3.1.

The fit Z_i and the residuals E_i of OLS are the (columnwise) orthogonal projections of Y_i on $\mathcal{R}(X_i)$ and $\mathcal{R}(X_i)^{\perp}$, respectively. In our specific EGLS, the fit $\widehat{\nu}_i$ is the orthogonal projection of Y_i on $\mathcal{R}(X_i\ \widehat{\varepsilon}_{(i-1)})$ with residuals $\widehat{\eta}_i \perp \mathcal{R}(X_i\ \widehat{\varepsilon}_{(i-1)})$. Figure 4.3.1 illustrates that Z_i and $\widehat{\mu}_i$ (and therefore E_i and $\widehat{\varepsilon}_i$) coincide when $\mathcal{R}(\widehat{\varepsilon}_{(i-1)}) \subseteq \mathcal{R}(X_i)^{\perp}$. So the equality $\widehat{\varepsilon}_i = E_i$ only holds if X_i and $\widehat{\varepsilon}_{(i-1)}$ are orthogonal; this is in general not the case.

We can distinguish several approaches for the construction of the EGLS estimator for Σ. First of all, it is possible to use the OLS estimator S, complete

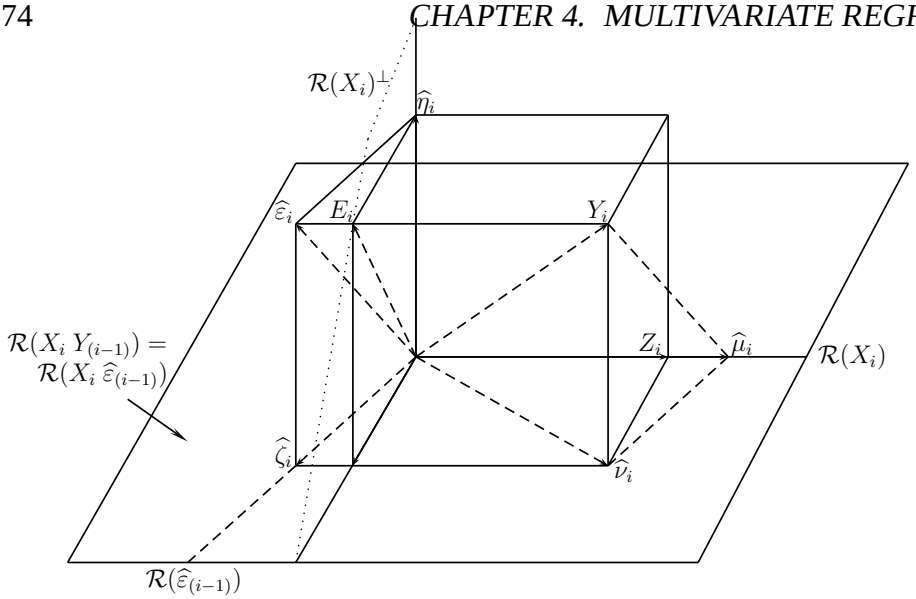


Figure 4.3.1: Geometric interpretation

ignoring all EGLS estimators. Secondly, it is also possible to base the EGLS estimator for Σ on the derived $\widehat{\varepsilon}_i$, while further ignoring $\widehat{\alpha}_i$. This approach is followed to construct the EGLS estimator in this section. Similar to OLS, we build the EGLS estimator $\widehat{S}$ for the covariance matrix as the sample variance corrected for degrees of freedom, *i.e.*

$$\begin{cases} \widehat{S}_{ii} = \widehat{\varepsilon}_i'\widehat{\varepsilon}_i/r_i \\ \widehat{S}_{ig} = \widehat{\varepsilon}_i'\widehat{\varepsilon}_{(i-1)g}/r_i \quad \text{for } g = 1, ..., i-1. \end{cases} \tag{4.3.21}$$

Similar to OLS, $\widehat{\varepsilon}_{(i-1)g}$ denotes the columns $M_{g-1}+1$ through M_g of $\widehat{\varepsilon}_{(i-1)}$, *i.e.* the first N_i rows of EGLS residuals corresponding to group g. Again, the estimator $\widehat{S}$ is not necessarily positive semidefinite. As a consequence of the regularity condition $N_r \geq M_r + l_r$, we have $r(\widehat{\varepsilon}_{(i-1)}) = M_{i-1}$ a.s., $l_{(i)} = l_i + M_{i-1}$ a.s. and the estimates $\widehat{\alpha}_i$ for the regression coefficients α_i are unique a.s..

Thirdly, and more logical, we could specify $\widehat{\Gamma}_{ii}$ since we already derived $\widehat{\alpha}_i$ (and (Γ_{ii}, α_i) completely specify Σ); we will discuss this approach in Section 4.3.5 in the context of ML.

4.3.5 Maximum likelihood

For ML estimation we make the additional assumption that the error terms ε_{ti} have (simultaneously for all t and i) a normal distribution. It follows that

$$\mathcal{L}\begin{pmatrix} Y_{t(i-1)} \\ Y_{ti} \end{pmatrix} = N_{M_i}\left(\begin{pmatrix} \mu_{t(i-1)} \\ \mu_{ti} \end{pmatrix}, \begin{pmatrix} \Sigma_{(i-1)(i-1)} & \Sigma_{(i-1)i} \\ \Sigma_{i(i-1)} & \Sigma_{ii} \end{pmatrix}\right).$$

The distribution of the observations is characterized by the unknown parameter $\theta = (\beta, \Sigma) \in \Theta$. We write $|A| = \det(A)$.

Theorem 4.3.4. *The likelihood of the observations $Y = \{Y_i\} = \{Y_{ti}\}$ is given by*

$$\begin{aligned} L(\theta; Y) &= \prod_{i=1}^{r} [\{(2\pi)^{m_i} |\Gamma_{ii}|\}^{-\frac{N_i}{2}} \exp\{-\tfrac{1}{2} tr(\Gamma_{ii}^{-1} \eta_i' \eta_i\}] && (4.3.22) \\ &= \exp\{-\tfrac{1}{2} \sum_{i=1}^{r} tr(\Gamma_{ii}^{-1} (\widehat{\nu}_i - \nu_i)'(\widehat{\nu}_i - \nu_i)\} \\ &\qquad \cdot \prod_{i=1}^{r} [\{(2\pi)^{m_i} |\Gamma_{ii}|\}^{-\frac{N_i}{2}} \exp\{-\tfrac{1}{2} tr(\Gamma_{ii}^{-1} \widehat{\eta}_i' \widehat{\eta}_i)\}]. && (4.3.23) \end{aligned}$$

Proof.

$$\begin{aligned} L(\theta; Y) &\overset{1}{=} \prod_{i=1}^{r} \prod_{t=1}^{N_i} p(Y_{ti} | Y_{t(i-1)}) \\ &\overset{2}{=} \prod_{i=1}^{r} [\{(2\pi)^{m_i} |\Gamma_{ii}|\}^{-\frac{N_i}{2}} \exp\{-\frac{1}{2} \sum_{t=1}^{N_i} (Y_{ti} - \nu_{ti})' \Gamma_{ii}^{-1} (Y_{ti} - \nu_{ti})\}] \\ &\overset{3}{=} \prod_{i=1}^{r} [\{(2\pi)^{m_i} |\Gamma_{ii}|\}^{-\frac{N_i}{2}} \exp\{-\frac{1}{2} tr(\Gamma_{ii}^{-1} (Y_i - \nu_i)'(Y_i - \nu_i)\}] \\ &\overset{4}{=} \prod_{i=1}^{r} [\{(2\pi)^{m_i} |\Gamma_{ii}|\}^{-\frac{N_i}{2}} \cdot \\ &\qquad \exp\{-\frac{1}{2} tr(\Gamma_{ii}^{-1} (\widehat{\nu}_i - \nu_i)'(\widehat{\nu}_i - \nu_i)) - \frac{1}{2} tr(\Gamma_{ii}^{-1} \widehat{\eta}_i' \widehat{\eta}_i)\}]. \end{aligned}$$

Equality 1 holds by conditioning; note that $Y_{t(0)} = 0$. Given $Y_{t(i-1)}$, ν_{ti} is fixed and (4.3.10) implies $\mathcal{L}(Y_{ti} | Y_{t(i-1)}) = N_{m_i}(\nu_{ti}, \Gamma_{ii})$. Because of the row independence the conditional densities can be substituted into the likelihood which results in

equality 2. Equality 3 is obtained by writing the likelihood in terms of matrices Y_i instead of the columns Y_{ti}; this proves (4.3.22). The fourth equality is based on the orthogonal decomposition of Y_i in $\widehat{\nu}_i$ and $\widehat{\eta}_i$ (according to (4.3.18)). Since $\widehat{\eta}_i$ is the orthogonal projection of Y_i onto $L_{(i)}^{\perp}$, $\widehat{\eta}_i$ is orthogonal to both $\widehat{\nu}_i$ and ν_i. This proves (4.3.23). □

In case of known Σ, it is clear from equality (4.3.22) that maximization of the likelihood coincides with minimization of the GLS criterion (4.3.13) and that the MLE's will coincide with the GLS estimators. So in case of normality, the GLS estimators are MVUE.

In case of unknown Σ, minimization of (4.3.23) leads to Theorem 4.3.5.

Theorem 4.3.5. *The MLE for μ_i coincides with the EGLS estimator $\widehat{\mu}_i$ as defined in Theorem 4.3.3. Moreover, the MLE for Γ_{ii} is*

$$\widehat{\Gamma}_{ii} = \frac{\widehat{\eta}_i'\widehat{\eta}_i}{N_i}. \tag{4.3.24}$$

The maximized likelihood is given by

$$\sup_{\vartheta\in\Theta} L(\theta; Y) = (2\pi e)^{-\frac{1}{2}\sum_{i=1}^{r} N_i m_i} \prod_{i=1}^{r} |\widehat{\eta}_i'\widehat{\eta}_i/N_i|^{-\frac{N_i}{2}}. \tag{4.3.25}$$

Proof. The MLE is obtained by maximization of the likelihood (4.3.23) w.r.t. all ν_i and Γ_{ii}, respectively. Now (4.3.23) is maximized by $\nu_i = \widehat{\nu}_i$, regardless the value of Γ_{ii}. Therefore $\widehat{\nu}_i$ is the MLE for ν_i, even in case of unknown Γ_{ii}. The estimators for the other parameters follow from (4.3.12) as in the case of EGLS estimation (see Section 4.3.4).

Substitution of $\widehat{\nu}_i$ in (4.3.23) gives

$$\sup_{\nu_i} L(\theta; Y) = \prod_{i=1}^{r} [\{(2\pi)^{m_i}|\Gamma_{ii}|\}^{-\frac{N_i}{2}} \exp\{-\tfrac{1}{2} tr(\Gamma_{ii}^{-1}\widehat{\eta}_i'\widehat{\eta}_i)\}].$$

This has to be maximized w.r.t. the Γ_{ii}. The separate factors of this maximized likelihood have the same structure as the expression for the complete multivariate linear model. So, in the same way we see that $\widehat{\Gamma}_{ii}$ of (4.3.24) is the MLE for Γ_{ii}. Substitution of the $\widehat{\nu}_i$ and $\widehat{\Gamma}_{ii}$ into (4.3.23) results in (4.3.25). □

In case of identifiable α_i and β_i, the EGLS estimators $\widehat{\beta}_i$ and $\widehat{\alpha}_i$ equal the MLE's. Though the coefficients α_i are identifiable, this is not true for β_i. In case of non-unique $\widehat{\beta}_i$ we choose the MLE equal to the EGLS estimator for β_i.

The MLE $\widehat{\Sigma}$ for the covariance matrix follows sequentially from the relations (4.3.8) and (4.3.9), and from the MLE (4.3.24):

$$\begin{cases} \widehat{\Sigma}_{11} = \widehat{\Gamma}_{11} \text{ and for } i = 2, ..., r: \\ \widehat{\Sigma}_{(i-1)i} = \widehat{\Sigma}_{(i-1)(i-1)}\widehat{\alpha}_i, \quad \widehat{\Delta}_{ii} = \widehat{\alpha}_i'\widehat{\Sigma}_{(i-1)(i-1)}\widehat{\alpha}_i, \quad \widehat{\Sigma}_{ii} = \widehat{\Gamma}_{ii} + \widehat{\Delta}_{ii}. \end{cases} \tag{4.3.26}$$

Note that the difference between the estimators $\widehat{\Sigma}$ and $\widehat{S}$ of (4.3.21) is not just caused by the introduction of the number of degrees of freedom. For example, from the expressions

$$\widehat{\Sigma}_{22} = \widehat{\eta}_2'\widehat{\eta}_2/N_2 + \widehat{\alpha}_2'\widehat{\varepsilon}_1'\widehat{\varepsilon}_1\widehat{\alpha}_2/N_1, \quad \widehat{S}_{22} = \widehat{\eta}_2'\widehat{\eta}_2/r_2 + \widehat{\alpha}_2'\widehat{\varepsilon}_{(1)}'\widehat{\varepsilon}_{(1)}\widehat{\alpha}_2/r_2$$

we see that the difference is caused by taking other residuals as well.

Note that we can use $\widehat{\Sigma}$ in EGLS (regardless of normality). It is not straightforward which one of the covariance matrix estimators S, $\widehat{S}$ or $\widehat{\Sigma}$ has the smallest bias. The bias of $\widehat{\Sigma}$ will probably be decreased by correcting for the degrees of freedom. Replacing N_i by $r_{(i)}$ in (4.3.24) gives an unbiased estimator for Γ_{ii}; Σ can still be estimated according to relation (4.3.26). A major drawback of this correction is that the estimator for Σ depends on the particular division of the data into groups, even in case of the complete model (with no missing observations). This problem is solved by substituting r_i for N_i in (4.3.24) and still estimating Σ by relation (4.3.26). Though this does not result in an unbiased estimator for Γ_{ii}, the estimator for Σ is unique in case of complete data and the bias of this estimator is probably smaller than the bias of the MLE $\widehat{\Sigma}$.

The analysis of the bias of the current covariance estimators S, $\widehat{S}$ and $\widehat{\Sigma}$ is left for future research. A similar approach as the one of Krishnamoorthy and Pannala (1999) or Kanda and Fujikoshi (1998) for the model with only the constant term could be followed. It would also be interesting to look at alternative estimators for the covariance matrix such as for example presented by Krishnamoorthy (1991) for the model with only the constant term. In this chapter, we restrict ourselves to (4.3.7), (4.3.21) and (4.3.26).

4.4 Relative efficiency

We compare the performance of the discussed LS estimators by means of the relative efficiency of the estimators for the regression coefficients under the normality assumption. The relative efficiency of estimator $\widehat{\theta}_1$ in relation to estimator $\widehat{\theta}_2$ can be expressed as the determinant of the following function of the M(ean) S(quared) E(rror)s:

$$MSE(\widehat{\theta}_1)^{-\frac{1}{2}} MSE(\widehat{\theta}_2) MSE(\widehat{\theta}_1)^{-\frac{1}{2}}, \tag{4.4.1}$$

other possibilities are the maximum eigenvalue or the trace.

Throughout this section we assume without loss of generality that $m_i = 1$ for all i. In case of normality all LS estimators for the regression coefficients are unbiased and their MSE's coincide with their variances. The variance of OLS estimator b_i follows directly from its definition in (4.3.5):

$$Var\{b_i\} = \sigma_{ii}(X_i'X_i)^{-1}. \tag{4.4.2}$$

The variance of the GLS estimator $\breve{\beta}_i$ is more complicated.

Theorem 4.4.1. *For $i = 2, \ldots, r$,*

$$Var\{\breve{\beta}_i\} = Var\{\breve{\beta}_{(i-1)}\alpha_i\} + (X_i'X_i)^{-1}X_i'\Gamma_{ii}X_i(X_i'X_i)^{-1}. \tag{4.4.3}$$

Proof. We determine the variance by the relation

$$Var\{\breve{\beta}_i\} = Var\{E\{\breve{\beta}_i|Y_{(i-1)}\}\} + E\{Var\{\breve{\beta}_i|Y_{(i-1)}\}\}.$$

For the variance of the conditional expectation we have

$$\begin{aligned} Var\{E\{\breve{\beta}_i|Y_{(i-1)}\}\} &\overset{1}{=} Var\{\beta_i + (X_i'X_i)^{-1}X_i'(\varepsilon_{(i-1)} - \breve{\varepsilon}_{(i-1)})\alpha_i\} \\ &\overset{2}{=} Var\{(X_i'X_i)^{-1}X_i'(X_i\breve{\beta}_{(i-1)} - X_i\beta_{(i-1)})\alpha_i\} \\ &\overset{3}{=} Var\{\breve{\beta}_{(i-1)}\alpha_i\}. \end{aligned}$$

The first equality follows from (4.3.16) and $E\{Y_i|Y_{(i-1)}\} = X_i\beta_i + \varepsilon_{(i-1)}\alpha_i$; the second from $\varepsilon_{(i-1)} - \breve{\varepsilon}_{(i-1)} = X_i\breve{\beta}_{(i-1)} - X_i\beta_{(i-1)}$ and $Var\{\beta_i\} = 0$. Rewriting and $Var\{\beta_{(i-1)}\} = 0$ gives the last equality.

For the conditional variance we have

$$\begin{aligned} Var\{\breve{\beta}_i|Y_{(i-1)}\} &= Var\{(X_i'X_i)^{-1}X_i'Y_i|Y_{(i-1)}\} \\ &= (X_i'X_i)^{-1}X_i'\Gamma_{ii}X_i(X_i'X_i)^{-1}, \end{aligned}$$

where the first equality follows from (4.3.16) and $Var\{\breve{\varepsilon}_{(i-1)}\alpha_i|Y_{(i-1)}\} = 0$; the second one from (4.3.12). □

Corollary. *If $M_2 = 2$, then*

$$Var(\breve{\beta}_2) = \rho_{12}^2\sigma_{22}(X_1'X_1)^{-1} + (1-\rho_{12}^2)\sigma_{22}(X_2'X_2)^{-1}, \tag{4.4.4}$$

where $\rho_{12} = \dfrac{\sigma_{12}}{\sqrt{\sigma_{11}\sigma_{22}}}$.

This corollary follows from Theorem 4.4.1, $\breve{\beta}_1 = b_1$ and (4.4.2).

We look into more detail at the relative efficiency for the frequently occurring situation $M_2 = 2$. Substituting (4.4.2) and (4.4.4) into (4.4.1) gives the relative efficiency of b_2 in relation to $\breve{\beta}_2$

$$(1-\rho_{12}^2) + \rho_{12}^2(X_2'X_2)^{\frac{1}{2}}(X_1'X_1)^{-1}(X_2'X_2)^{\frac{1}{2}}. \tag{4.4.5}$$

It is clear that (4.4.5) is always smaller (or equal) to one, *i.e.* $\breve{\beta}_2$ always outperforms b_2 in terms of variance (as can be expected). GLS is relatively more efficient for high values of ρ_{12} and small $(X_2'X_2)(X_1'X_1)^{-1}$; the latter usually corresponds with a high fraction of missing observations, *i.e.* n_2/n_1 is small. This seems to be quite a logical result: GLS makes use of the sample information of preceding dependent variables in contrast to OLS. If there is relatively a lot of additional information available (*i.e.* n_1/n_2) is high) and the preceding dependent variable is highly correlated with the current one, the additional information concerning the preceding dependent variable will result in more accurate estimates. Figure 4.4.1 plots the relative efficiency of b_2 in relation to $\breve{\beta}_2$ as function of ρ_{12} for

several combinations of n_1/n_2 (under the assumption that $(X_2'X_2)(X_1'X_1)^{-1}$ is equivalent to n_2/n_1).

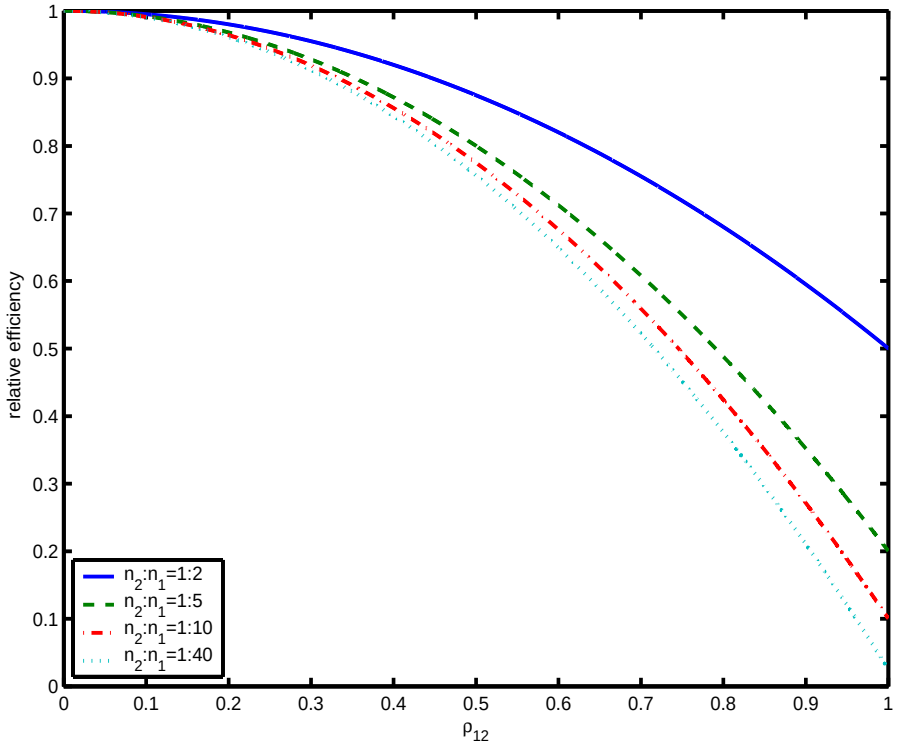


Figure 4.4.1: Relative efficiency of b_2 in relation to $\tilde{\beta}_2$

It is quite hard to derive a closed form expression for $Var\{\widehat{\beta}_i\}$. However, (4.4.3) will give a good approximation for large sample sizes since EGLS is asymptotically equivalent to GLS. In Chapter 6 we will consider the relative efficiency of OLS in relation to EGLS for a practical example.

4.5 Special cases

4.5.1 No missing observations

In the model formulation of Section 4.2 the restrictions $N_{i-1} \geq N_i$ are imposed instead of $N_{i-1} > N_i$. In case of the last restrictions the division of the data into

different group is always unique, while this is not true for the first restrictions: if there are several variables with the same number of observations, all the variables together can be defined as one group, but it is also possible to define multiple groups. In case of different groups with the same number of observations, the $\widehat{\varepsilon}_j$ of the previous dependent variables with the same number of observations as the dependent variables of group i are orthogonal to X_i. Since the regression of Y_i onto the X_i and $\widehat{\varepsilon}_{(i-1)}$ coincides with partial regression (see *e.g.* Green (1993)), the estimators $\widehat{\mu}_i$ and $\widehat{\varepsilon}_i$ will not depend on the group composition.

The situation with no missing observations ($N = N_1 = ... = N_r$) is a special case of the presented model. By constructing just one group, it is straightforward that the OLS and (E)GLS estimators for μ_i are identical: $Z_i = \widehat{\mu}_i$. As a consequence the covariance estimators (4.3.7) and (4.3.21) are identical and unique.

The uniqueness and equality of the OLS and EGLS estimators can also be shown sequentially by the estimation procedure. From both Figure 4.3.1 and formula (4.3.20) for the regression coefficients, we can see that the OLS and EGLS estimators are identical when $\mathcal{R}(\widehat{\varepsilon}_{(i-1)}) \subseteq \mathcal{R}(X_i)^\perp$. That this is true for the situation with no missing observations can be directly deduced from the estimation procedure. In case of complete data, we have $X = X_1 = X_2 = ... = X_r$ and $\mathcal{R}(X_i) = \mathcal{R}(X)$ for $i = 1, ..., r$. The iterations in the EGLS estimation procedure show

$$\begin{array}{lll}
\text{Step 1:} & & \widehat{\mu}_1 \in \mathcal{R}(X), \quad \widehat{\varepsilon}_1 \in \mathcal{R}(X)^\perp \\
\text{Step } i \ (i = 2, ..., r): & & \widehat{\varepsilon}_{(i-1)} = [\widehat{\varepsilon}_1 \ \widehat{\varepsilon}_2 \ ... \ \widehat{\varepsilon}_{i-1}] \in \mathcal{R}(X)^\perp \\
& \Longrightarrow & \widehat{\zeta}_i = \widehat{\varepsilon}_{(i-1)}\widehat{\alpha}_i \in \mathcal{R}(\widehat{\varepsilon}_{(i-1)}) \subseteq \mathcal{R}(X)^\perp \text{ and } \widehat{\eta}_i \in \mathcal{R}(X)^\perp \\
& \Longrightarrow & \widehat{\varepsilon}_i = \widehat{\zeta}_i + \widehat{\eta}_i \in \mathcal{R}(X)^\perp \\
& \Longrightarrow & \mathcal{R}(\widehat{\varepsilon}_{(i)}) = \mathcal{R}(\widehat{\varepsilon}_1 \ \widehat{\varepsilon}_2 \ ... \ \widehat{\varepsilon}_i) \subseteq \mathcal{R}(X)^\perp.
\end{array}$$

So $\widehat{\varepsilon}_{(i-1)} \in \mathcal{R}(X)^\perp$, $Z_i = \widehat{\mu}_i$ and as a consequence $S = \widehat{S}$.

For the case of complete data, the MLE in Theorem 4.3.5 must be identical to the standard result known from literature, as well as the maximized likelihood (4.3.25). To show the latter we make use of the following two properties:

(a) $\widehat{\eta}_1, \widehat{\eta}_2, \ldots, \widehat{\eta}_r$ are orthogonal,
(b) $E_{(r)} = [E_1 \quad E_2 \ldots E_r] = [\widehat{\varepsilon}_1 \quad \widehat{\varepsilon}_2 \ldots \widehat{\varepsilon}_r] = [\widehat{\eta}_1 \quad \widehat{\eta}_2 \ldots \widehat{\eta}_r]A$.

with A an upper triangular invertible matrix with unit diagonal elements. The existence of such a matrix follows form (a slightly modified) QR-factorization of $E_{(r)}$. Hence,

$$\begin{aligned}
\sup_{\vartheta\in\Theta} L(\theta;Y) &\stackrel{1}{=} (2\pi e)^{-\frac{1}{2}NM} \prod_{i=1}^{r} |\widehat{\eta}_i'\widehat{\eta}_i/N|^{-\frac{N}{2}} \\
&\stackrel{2}{=} (2\pi e)^{-\frac{1}{2}NM} |[\widehat{\eta}_1 \quad \widehat{\eta}_2 \ldots \widehat{\eta}_r]'[\widehat{\eta}_1 \quad \widehat{\eta}_2 \ldots \widehat{\eta}_r]/N|^{-\frac{N}{2}} \\
&\stackrel{3}{=} (2\pi e)^{-\frac{1}{2}NM} |[\widehat{\eta}_1 \quad \widehat{\eta}_2 \ldots \widehat{\eta}_r]A)'([\widehat{\eta}_1 \quad \widehat{\eta}_2 \ldots \widehat{\eta}_r]A/N|^{-\frac{N}{2}} \\
&\stackrel{4}{=} (2\pi e)^{-\frac{1}{2}NM} |E_{(r)}'E_{(r)}/N|^{-\frac{N}{2}}.
\end{aligned}$$

The first equality follows from $N = N_1 = ... = N_r$ and $X = X_1 = ... = X_r$. The second and third equality are based on property (a) and $|A| = 1$. The last equality follows from (b). The final expression can be found in Seber (1984), p. 407. A general approach for complete data can be found in Van der Genugten (1997) *e.g.*, emphasizing a geometrical approach.

4.5.2 The constant term

Often the first explanatory variable is the constant term. We denote the corresponding regression coefficients by $\beta_c \in I\!R^{1\times M}$ (c = constant); the regression coefficients of the other explanatory variables are denoted by $\beta_v \in I\!R^{(k-1)\times M}$ (v = variable). Expanding this notation we can write

$$\beta = \begin{bmatrix} \beta_c \\ \beta_v \end{bmatrix}, \quad b = \begin{bmatrix} b_c \\ b_v \end{bmatrix}, \quad \widehat{\beta} = \begin{bmatrix} \widehat{\beta}_c \\ \widehat{\beta}_v \end{bmatrix}, \quad X = [1_N \; X_v],$$

with $X_v \in I\!R^{N_1\times(k-1)}$. The subindices i and $(i-1)$ have a similar meaning as in the preceding sections, so for example, X_{vi} contains the first N_i rows of X_v.

LS estimation with the constant term corresponds to orthogonal projections on $\mathcal{R}(1_{N_i})$ and the centered spaces $\widetilde{L}_i$ and $\widetilde{L}_{(i)}$ defined as

$$\begin{array}{llll}
\widetilde{L}_i \oplus \mathcal{R}(1_{N_i}) = L_i & \text{and} & \widetilde{L}_i \perp \mathcal{R}(1_{N_i}), & \widetilde{l}_i = \dim(\widetilde{L}_i) = l_i - 1, \\
\widetilde{L}_{(i)} \oplus \mathcal{R}(1_{N_i}) = L_{(i)} & \text{and} & \widetilde{L}_{(i)} \perp \mathcal{R}(1_{N_i}), & \widetilde{l}_{(i)} = \dim(\widetilde{L}_{(i)}) = l_{(i)} - 1.
\end{array}$$

The mean and centered observations coincide with orthogonal projections of the

observations on $\mathcal{R}(1_{N_i})$ and the centered spaces:

$$\begin{array}{ll}
\overline{X}_i = \frac{1}{N_i} 1'_{N_i} X_{vi} \in I\!R^{1\times(k-1)}, & \widetilde{X}_i = X_{vi} - 1_{N_i}\overline{X}_i \in I\!R^{N_i\times(k-1)}, \\
\overline{Y}_i = \frac{1}{N_i} 1'_{N_i} Y_i \in I\!R^{1\times m_i}, & \widetilde{Y}_i = Y_i - 1_{N_i}\overline{Y}_i \in I\!R^{N_i\times m_i}, \\
\overline{Y}_{(i-1)} = \frac{1}{N_i} 1'_{N_i} Y_{(i-1)} \in I\!R^{1\times M_{i-1}}, & \widetilde{Y}_{(i-1)} = Y_{(i-1)} - 1_{N_i}\overline{Y}_{(i-1)} \in I\!R^{N_i\times M_{i-1}}, \\
\overline{\varepsilon}_{(i-1)} = \frac{1}{N_i} 1'_{N_i} \widehat{\varepsilon}_{(i-1)} \in I\!R^{1\times M_{i-1}}, & \widetilde{\varepsilon}_{(i-1)} = \widehat{\varepsilon}_{(i-1)} - 1_{N_i}\overline{\varepsilon}_{(i-1)} \in I\!R^{N_i\times M_{i-1}}.
\end{array}$$

Note that $\overline{Y}_{(i-1)} \neq [\overline{Y}_1\ \overline{Y}_2 \ldots \overline{Y}_{i-1}]$ and $\overline{\varepsilon}_{(i-1)} \neq 0$.

The LS estimators can be expressed in terms of the means and the centered observations, *e.g.* the EGLS estimators (or equivalently the MLE's in case of normality and unknown Σ) read

$$\left\{ \begin{array}{l}
\left[\begin{array}{c} \widehat{\beta}_{vi} \\ \widehat{\alpha}_i \end{array} \right] = \widetilde{G}_{(i)} \left[\begin{array}{c} \widetilde{X}'_i \\ \widetilde{\varepsilon}'_{(i-1)} \end{array} \right] \widetilde{Y}_i, \quad \text{with } \widetilde{G}_{(i)} = \left[\begin{array}{cc} \widetilde{X}'_i\widetilde{X}_i & \widetilde{X}'_i\widetilde{\varepsilon}_{(i-1)} \\ \widetilde{\varepsilon}'_{(i-1)}\widetilde{X}_i & \widetilde{\varepsilon}'_{(i-1)}\widetilde{\varepsilon}_{(i-1)} \end{array} \right]^- \\
\\
\widehat{\beta}_{ci} = \overline{Y}_i - \overline{X}_i\widehat{\beta}_{vi} - \overline{\varepsilon}_{(i-1)}\widehat{\alpha}_i.
\end{array} \right. \tag{4.5.1}$$

We now turn to the very special case that the constant term is the only explanatory variable: $X_i = 1_{N_i}$. This model has received considerable attention in literature, especially ML estimation under the normality assumption. Anderson (1957) derived the MLE's for $r = 2$ and $m_1 = m_2 = 1$ and suggested an approach to determine the MLE's for general r. Bhargava (1962) derived the MLE's for general r. Following the approach suggested by Anderson (1957), Afifi and Elashoff (1966) confirmed the findings of Bhargava (1962) for the regression coefficients, but presented a different, incorrect MLE for the covariance matrix. Jinadasa and Tracy (1992) derived the correct MLE's for general r by matrix differentiation which resulted in rather complicated expressions. Fujisawa (1995) presented the MLE's for general r in recursive form, which coincide with the MLE's given by Bhargava (1962) and Jinadasa and Tracy (1992).

For the model with only the constant term, $\widetilde{\varepsilon}_{(i-1)}$ and $\widetilde{\varepsilon}_i$ coincide with $\widetilde{Y}_{(i-1)}$ and $\widetilde{Y}_i$ respectively, and the MLE's (4.5.1) for the regression coefficients reduce

to the same expressions as found by Fujisawa (1995):

$$\begin{array}{l}\widehat{\beta}_{c1} = \widehat{\mu}_1 = \overline{Y}_1, \\ \widehat{\alpha}_i = (\widetilde{Y}'_{(i-1)}\widetilde{Y}_{(i-1)})^{-1}(\widetilde{Y}_{(i-1)}\widetilde{Y}_i), \\ \widehat{\beta}_{ci} = \widehat{\mu}_i = \overline{Y}_i - (\overline{Y}_{(i-1)} - \widehat{\mu}_{(i-1)})\widehat{\alpha}_i, \quad \text{for } i = 2, ..., r.\end{array}$$

The MLE $\widehat{\Gamma}_{ii}$ is determined by substituting the MLE's for the regression coefficients into (4.3.24), leading to the same covariance estimators as found by Fujisawa (1995):

$$\left\{\begin{array}{l}\widehat{\Gamma}_{11} = \widetilde{Y}_1'\widetilde{Y}_1/N_1 \\ \widehat{\Gamma}_{ii} = (\widetilde{Y}_i - \widetilde{Y}_{(i-1)}\widehat{\alpha}_i)'(\widetilde{Y}_i - \widetilde{Y}_{(i-1)}\widehat{\alpha}_i)/N_i \quad \text{for } i = 2, ..., r.\end{array}\right.$$

4.6 Restricted models

So far we just have considered (unrestricted) models in which $\mu_i \in L_i$ and $\nu_i \in L_{(i)}$. In a restricted model, p_i linear constraints are imposed on the parameters β_i : $C_i\beta_i = 0$ with $C_i \in I\!R^{p_i \times k}$ for $i = 1, \ldots, r$. So for $i = 1, \ldots, r$ the unknown β_i are restricted to $\mathcal{N}(C_i)$, the null space of C_i. We assume that the restrictions are monotone (decreasing) in the sense that $\mathcal{N}(C_1) \subseteq \mathcal{N}(C_2) \subseteq \ldots \subseteq \mathcal{N}(C_r)$. This includes the usual case $C_1 = \ldots = C_r$.

Similar to the unrestricted model, we can distinguish between OLS , (E)GLS and ML estimation. We will only discuss the specific EGLS corresponding to ML under normality.

For two matrices $P \in I\!R^{p\times c}$and $Q \in I\!R^{q\times c}$ we will write $\begin{bmatrix} P \\ Q \end{bmatrix}$ shortly as $[P; Q]$. Now $\nu_i = [X_i\ Y_{(i-1)}][\beta_i; \alpha_i]$ is restricted to $\mathcal{R}(X_i(\mathcal{N}(C_i))\ Y_{(i-1)})$, where $X_i(\mathcal{N}(C_i))$ is the image of $\mathcal{N}(C_i)$ under the linear transformation X_i. The linear space $L_{(i)} = \mathcal{R}(X_i\ Y_{(i-1)})$ can be split into two orthogonal subspaces: $L_{0(i)}$ and $L_{1(i)}$, which (with some additional characteristics) are defined as

$$\begin{array}{ll}L_{0(i)} = \mathcal{R}(X_i(\mathcal{N}(C_i))\ Y_{(i-1)}), & L_{1(i)} \oplus L_{0(i)} = L_{(i)}, \quad L_{1(i)} \perp L_{0(i)}, \\ H_{0(i)}\text{: projection matrix of } L_{0(i)}, & l_{0(i)} = \dim(L_{0(i)}), \\ H_{1(i)}\text{: projection matrix of } L_{1(i)}, & l_{1(i)} = \dim(L_{1(i)}) = l_{(i)} - l_{0(i)}, \\ U_{0(i)}\text{: projection matrix of } L^{\perp}_{0(i)}, & L_{0(i)} \oplus L^{\perp}_{0(i)} = I\!R^{N_i}, \quad L_{0(i)} \perp L^{\perp}_{0(i)}, \\ & r_{0(i)} = \dim(L^{\perp}_{0(i)}) = N_i - l_{0(i)}.\end{array}$$

So $L_{0(i)}^{\perp} = L_{1(i)} \oplus L_{(i)}^{\perp}$. Quantities relating to $L_{0(i)}$ and $L_{1(i)}$ are denoted by a primary subindex 0 and 1, respectively. The following testing problem will be considered (for identifiable $C_i\beta_i$):

$$\left\{ \begin{array}{l} H_0 : \{\forall i : C_i\beta_i = 0\} \text{ against } H_1 : \{\exists i : C_i\beta_i \neq 0\}, \\ \quad \text{or equivelantly,} \\ H_0 : \{\forall i : \nu_i \in L_{0(i)}\} \text{ against } H_1 : \{\exists i : \nu_i \in L_{(i)} - L_{0(i)}; \forall i : \nu_i \in L_{(i)}\}. \end{array} \right. \tag{4.6.1}$$

The relevant test statistics for (4.6.1) can be based on orthogonal projections onto the $L_{1(i)}$ and $L_{(i)}^{\perp}$.

The whole procedure for EGLS estimation for the restricted model is similar to the one described in Section 4.3.4 for the unrestricted model: only the subspaces $L_{(i)}$ have to be replaced by $L_{0(i)}$. This is due to the fact that the restrictions are monotone, implying that $\mu_i, \mu_{(i-1)} \in L_{0(i)}$. Formulae (4.3.18), (4.3.19) and (4.3.21) through (4.3.26) still hold for the restricted model if we add a subindex 0. The estimators $\widehat{\beta}_{0i}$ and $\widehat{\alpha}_{0i}$ for β_i and α_i respectively, are given (similar to (4.3.20)) by

$$\left\{ \begin{array}{l} \widehat{\beta}_{01} = G_{01} X_1' Y_1, \text{ with } G_{01} \in I\!R^{k \times k} \\ \text{defined by } \left[\begin{array}{cc} G_{01} & * \\ * & * \end{array} \right] = \left[\begin{array}{c:c} X_1'X_1 & C_1' \\ \hdashline C_1 & 0 \end{array} \right]^{-} \\ i = 2, \ldots, r : \\ \quad \left[\begin{array}{c} \widehat{\beta}_{0i} \\ \widehat{\alpha}_{0i} \end{array} \right] = G_{0(i)} \left[\begin{array}{c} X_i' \\ \widehat{\varepsilon}_{0(i-1)}' \end{array} \right] Y_i \text{ with } G_{0(i)} \in I\!R^{(k+M_{i-1}) \times (k+M_{i-1})} \\ \quad \text{defined by } \left[\begin{array}{cc} G_{0(i)} & * \\ * & * \end{array} \right] = \left[\begin{array}{c:cc} X_i'X_i & X_i'\widehat{\varepsilon}_{0(i-1)} & C_i' \\ \widehat{\varepsilon}_{0(i-1)}'X_i & \widehat{\varepsilon}_{0(i-1)}'\widehat{\varepsilon}_{0(i-1)} & 0 \\ \hdashline C_i & 0 & 0 \end{array} \right]^{-}. \end{array} \right. \tag{4.6.2}$$

The required statistics for the LR test (based on EGLS) can be summarized into a collection of non-centered MANOVA-tables for $i = 1, \ldots, r$. In the tables the abbreviations SS, DF and R stand for Sum of Squares, Degrees of Freedom and Restricted, respectively.

Model	Space	SS	DF	Testing
R. model	$L_{0(i)}$	$\widehat{\nu}'_{0i}\widehat{\nu}_{0i}$	$l_{0(i)}$	$\Lambda_{0i} = \dfrac{\lvert\widehat{\eta}'_i\widehat{\eta}_i\rvert}{\lvert\widehat{\eta}'_i\widehat{\eta}_i + \widehat{\nu}'_{1i}\widehat{\nu}_{1i}\rvert}$
Difference	$L_{1(i)}$	$\widehat{\nu}'_{1i}\widehat{\nu}_{1i}$	$l_{1(i)}$	
Model	$L_{(i)}$	$\widehat{\nu}'_i\widehat{\nu}_i$	$l_{(i)}$	
Error	$L^{\perp}_{(i)}$	$\widehat{\eta}'_i\widehat{\eta}_i$	$r_{(i)}$	
Total	$\mathbb{R}^{N_i}$	Y'_iY_i	N_i	

Table 4.6.1: Collection of non-centered MANOVA-tables $(i = 2, \ldots, r)$

The column Testing will be used in case of normality in Section 4.8; note that $\widehat{\eta}_{0i} = U_{0(i)}Y_i = \widehat{\nu}_{1i} + \widehat{\eta}_i$.

If the constant term is included as an explanatory variable, often the centered MANOVA-tables are presented, provided that no restrictions are imposed on the constant term. The abbreviation C stands for Corrected (or Centered):

Model	Space	SS	DF	Testing
C.R. model	$\widetilde{L}_{0(i)}$	$\widetilde{\nu}'_{0i}\widetilde{\nu}_{0i}$	$\widetilde{l}_{0(i)}$	$\Lambda_{0i} = \dfrac{\lvert\widehat{\eta}'_i\widehat{\eta}_i\rvert}{\lvert\widehat{\eta}'_i\widehat{\eta}_i + \widehat{\nu}'_{1i}\widehat{\nu}_{1i}\rvert}$
Difference	$L_{1(i)}$	$\widehat{\nu}'_{1i}\widehat{\nu}_{1i}$	$l_{1(i)}$	
C. model	$\widetilde{L}_{(i)}$	$\widetilde{\nu}'_i\widetilde{\nu}_i$	$\widetilde{l}_{(i)}$	
Error	$L^{\perp}_{(i)}$	$\widehat{\eta}'_i\widehat{\eta}_i$	$r_{(i)}$	
C. total	$\mathcal{R}(1_{N_i})^{\perp}$	$\widetilde{Y}'_i\widetilde{Y}_i$	$N_i - 1$	
Mean	$\mathcal{R}(1_{N_i})$	$N_i\overline{Y}'_i\overline{Y}_i$	1	
Total	$\mathbb{R}^{N_i}$	Y'_iY_i	N_i	

Table 4.6.2: Collection of centered MANOVA-tables $(i = 2, \ldots, r)$

The inner products in the non-centered MANOVA-tables are acquired by adding the inner products of the corresponding means to the centered inner products, *e.g.* $\widehat{\nu}'_i\widehat{\nu}_i = \widetilde{\nu}'_i\widetilde{\nu}_i + N_i\overline{Y}'_i\overline{Y}_i$. Since the terms $\widehat{\nu}_{1i}$ and the errors $\widehat{\eta}_i$ in the non-centered MANOVA-tables are centered if a constant is included in the model, they are identical to the corresponding inner products in the centered MANOVA-tables.

Now suppose that (not necessary identifiable) linear restrictions $C_i\beta_i = 0$ have already been imposed and that q_i additional linear constraints are considered of the form $D_i\beta_i = 0$ with $D_i \in \mathbb{R}^{q_i \times k}$. Then the unknown β_i is restricted

to $\mathcal{N}([C_i; D_i])$, the null space of $[C_i; D_i]$. This double restricted model (with $[C_i; D_i]\beta_i = 0$) is discussed here, since this model enables us to formulate and solve the most general case; there is no need for additional triple constraints.

Again, we assume that the additional restrictions are monotone: $\mathcal{N}(D_1) \subseteq \mathcal{N}(D_2) \subseteq \ldots \subseteq \mathcal{N}(D_r)$. Similar to the (single) restricted model, the linear space $L_{0(i)}$ can be split into the subspaces $L_{00(i)} = \mathcal{R}(X_i(\mathcal{N}([C_i; D_i]))\ Y_{(i-1)})$ and $L_{01(i)}$, the orthogonal complement of $L_{00(i)}$ w.r.t. $L_{0(i)}$. We will consider the following testing problem (for identifiable $[C_i; D_i]\beta_i$):

$$\left\{\begin{array}{l} H_{00} : \{\forall i : C_i\beta_i = 0, D_i\beta_i = 0\} \text{ against } H_{01} : \{\exists i : D_i\beta_i \neq 0; \forall i : C_i\beta_i = 0\} \\ \quad \text{or equivalently,} \\ H_{00} : \{\forall i : \nu_i \in L_{00(i)}\} \text{ against } H_{01} : \{\exists i : \nu_i \in L_{0(i)} - L_{00(i)}; \forall i : \nu_i \in L_{0(i)}\} \end{array}\right. \tag{4.6.3}$$

The test statistics for (4.6.3) can be based on orthogonal projections onto the $L_{01(i)}$ and $L_{0(i)}^{\perp}$. The estimation procedure of the preceding sections can again be applied to the double restricted model similar as to the restricted model. For estimation under the (not necessarily identifiable) double restrictions $[C_i; D_i]\beta_i = 0\ \forall i$, we can use again (4.6.2) with C_i replaced by $[C_i; D_i]$.

All information of the unrestricted, restricted and double restricted models required for the described tests can be summarized in combined centered MANOVA-tables for $i = 1, \ldots, r$, assuming that the model contains the constant as an explanatory variable and that no restrictions are imposed on this constant. This combined centered MANOVA-table can be obtained by adding Table 4.6.3 to the top of the centered MANOVA-table in Table 4.6.2. Here D stands for double:

Model	**Space**	**SS**	**DF**	**Testing**
C. D. Restricted model	$\widetilde{L}_{00(i)}$	$\widetilde{\nu}'_{00i}\widetilde{\nu}_{00i}$	$\widetilde{l}_{00(i)}$	
Difference	$L_{01(i)}$	$\widehat{\nu}'_{01i}\widehat{\nu}_{01i}$	$l_{01(i)}$	$\Lambda_{00i} = \dfrac{\lvert\widehat{\eta}'_{0i}\widehat{\eta}_{0i}\rvert}{\lvert\widehat{\eta}'_{0i}\widehat{\eta}_{0i} + \widehat{\nu}'_{01i}\widehat{\nu}_{01i}\rvert}$

Table 4.6.3: Double restricted centered inner products $(i = 2, \ldots, r)$

From Tables 4.6.1, 4.6.2 and 4.6.3 relations between the unrestricted, restricted and double restricted statistics can be deduced such as $\widehat{\eta}'_{00i}\widehat{\eta}_{00i} = \widehat{\eta}'_{0i}\widehat{\eta}_{0i} + \widehat{\nu}'_{01i}\widehat{\nu}_{01i}$.

The related testing procedure will be discussed for the normal case in Section 4.8.

4.7 Some distributions and orthogonal projections

We define the Wishart distribution W_d as follows: let $Y = [Y_1 \dots Y_n]'$ and $\mu = [\mu_1 \dots \mu_n]'$ with independent $Y_i \sim N_d(\mu_i, \Sigma), \Sigma \geq 0$. Then

$$W = Y'Y = \sum_{i=1}^{n} Y_i Y_i' \sim W_d(n, \Sigma; \Delta) \quad \text{(with } \Delta = \mu'\mu),$$

where $W_d(n, \Sigma; \Delta)$ denotes the noncentral Wishart distribution with dimension d, degrees of freedom n, dispersion matrix Σ and non-centrality matrix Δ. The central Wishart distribution is $W_d(n, \Sigma) = W_d(n, \Sigma; 0)$. The standard Wishart distribution is $W_d(n) = W_d(n, I_d)$. Our notation is the same as the one of Gupta and Nagar (2000), except for the non-centrality matrix which they define as $\Theta = \Sigma^{-1}\Delta$ for $\Sigma > 0$. We prefer to include singular Σ as well.

The properties of the projections follow from the following projection theorem (compare Gupta and Nagar (2000), Theorems 7.8.3 and 7.8.5).

Theorem 4.7.1. *Let L_0 and L_1 be linear subspaces of $\mathbb{R}^n$ with $L_0 \perp L_1$. Denote the orthogonal projection matrices of L_0 and L_1 by P_0 respectively P_1 and let $l_0 = dim(L_0)$. Then, for $Y' = [Y_1 \dots Y_n] \in \mathbb{R}^{d \times n}$, with uncorrelated Y_i, $Cov(Y_i) = \Sigma$ and $E\{Y\} = \mu$,*

P_0Y and P_1Y are uncorrelated,
$E\{P_0Y\} = P_0\mu$,
$Cov(vec(P_0Y)) = \Sigma \otimes P_0$.

If in addition the Y_i are normally distributed, then

P_0Y and P_1Y are independent,
$Y'P_0Y \sim W_d(l_0, \Sigma; \mu'P_0\mu)$.

In the next section a generalization of the Wilks' distribution is used. For the (usual) Wilks' distribution we follow the same notation as *e.g.* Rencher (1998): let $B \sim W_d(s, \Sigma)$, $C \sim W_d(t, \Sigma)$, B and C independent. Then

$$\Lambda = \frac{|B|}{|B + C|} \sim \Lambda_{d,t,s} \quad ,$$

where $\Lambda_{d,t,s}$ denotes the Wilks' distribution with parameters d, t and s. We define the generalized Wilks' distribution $\Lambda_{A,D,T,S}$ with parameter vectors A, D, T and $S \in I\!R^{1\times r}$as follows: let $\Lambda_i \sim \Lambda_{d_i,t_i,s_i}$ be independent and $a_i \in [0,1]$ with $a_1 = 1$. Then, by definition,

$$\prod_{i=1}^{r} \Lambda_i^{a_i} \sim \Lambda_{A,D,T,S}.$$

The vector A contains the exponents a_i of the separate factors as elements, D the d_i, T the t_i and S the s_i $(i = 1, \ldots, r)$.

4.8 Testing

We assume normally distributed errors now. From the projection Theorem 4.7.1 (applied to $L_{(i)}$ and $L_{(i)}^{\perp}$) we get the following conditional properties given $Y_{(i-1)}$:

$$\left\{\begin{array}{ll} \widehat{\nu}_i \text{ and } \widehat{\eta}_i \text{ are independent, normally distributed conditional under } Y_{(i-1)} & \\ \mathrm{E}\{\widehat{\nu}_i|Y_{(i-1)}\} = H_{(i)}\nu_i = \nu_i, & \mathrm{E}\{\widehat{\eta}_i|Y_{(i-1)}\} = U_{(i)}\nu_i = 0 \\ \mathrm{Cov}(\mathrm{vec}(\widehat{\nu}_i)|Y_{(i-1)}) = \Gamma_{ii} \otimes H_{(i)}, & \mathrm{Cov}(\mathrm{vec}(\widehat{\eta}_i)|Y_{(i-1)}) = \Gamma_{ii} \otimes U_{(i)} \\ \mathcal{L}(\widehat{\nu}_i'\widehat{\nu}_i|Y_{(i-1)}) = W_{m_i}(l_{(i)}, \Gamma_{ii}; \nu_i'\nu_i), & \mathcal{L}(\widehat{\eta}_i'\widehat{\eta}_i|Y_{(i-1)}) = W_{m_i}(r_{(i)}, \Gamma_{ii}). \end{array}\right. \tag{4.8.1}$$

Here we have used that $\nu_i' H_{(i)} \nu_i = \nu_i'\nu_i$ and $\nu_i' U_{(i)} \nu_i = 0$. These properties permit us to give confidence intervals for (identifiable) $C_i\beta_i$. We omit the details and concentrate on testing.

The following unconditional properties also hold

$$\left\{\begin{array}{l} Y_{(i-1)}, \widehat{\nu}_i \text{ and } \widehat{\eta}_i \text{ are normally distributed} \\ \mathcal{L}(\widehat{\eta}_i'\widehat{\eta}_i) = W_{m_i}(r_{(i)}, \Gamma_{ii}) \\ (Y_{(i-1)}, \widehat{\nu}_i \text{) and } \widehat{\eta}_i'\widehat{\eta}_i \text{ are independent} \\ \widehat{\nu}_1, \widehat{\eta}_1, \widehat{\eta}_2'\widehat{\eta}_2, ..., \widehat{\eta}_r'\widehat{\eta}_r \text{ are independent.} \end{array}\right. \tag{4.8.2}$$

The first three properties follow directly from (4.8.1); the last from the fact that $\widehat{\eta}_j$ $(j < i)$ is a function of $Y_{(j-1)}$ and Y_j and therefore of $Y_{(i-1)}$ and the individual observations $Y_{t(j)}$, $t = N_i + 1, \ldots, N_j$. The latter are independent of $\widehat{\eta}_i$ because of the row independence of the observations (see (4.2.2)).

Now consider the likelihood ratio test for the hypothesis (4.6.1). Denote the restricted parameter space of $\theta = (\beta, \Sigma)$ by Θ_0. From (4.3.25) the likelihood ratio LR_0 for (4.6.1) is given by

$$\begin{aligned} LR_0 &= \frac{\sup_{\theta\in\Theta_0} L(\theta; Y)}{\sup_{\theta\in\Theta} L(\theta; Y)} = \prod_{i=1}^{r} \left(\frac{|\widehat{\eta}_i'\widehat{\eta}_i|}{|\widehat{\eta}_{0i}'\widehat{\eta}_{0i}|} \right)^{\frac{N_i}{2}} = \prod_{i=1}^{r} \left(\frac{|\widehat{\eta}_i'\widehat{\eta}_i|}{|\widehat{\eta}_i'\widehat{\eta}_i + \widehat{\nu}_{1i}'\widehat{\nu}_{1i}|} \right)^{\frac{N_i}{2}} \\ &= \prod_{i=1}^{r} \Lambda_{0i}^{\frac{N_i}{2}}. \end{aligned} \tag{4.8.3}$$

For the model with only the constant term as explanatory variable, LR_0 reduces to the test statistic which Bhargava (1962) derived. Hao and Krishnamoorthy (2001) discussed that test statistic in more detail; in both papers its distribution was approximated.

Since $\nu_i' H_{1(i)} \nu_i = 0$ under H_0 of (4.6.1), applying Theorem 4.7.1 to $\widehat{\nu}_{0i} \in L_{0(i)}$, $\widehat{\nu}_{1i} \in L_{1(i)}$ and $\widehat{\eta}_i \in L_{(i)}^{\perp}$ leads to the conclusion that $\widehat{\nu}_{11}'\widehat{\nu}_{11}, \widehat{\nu}_{12}'\widehat{\nu}_{12}, \ldots, \widehat{\nu}_{1r}'\widehat{\nu}_{1r}$, $\widehat{\eta}_1'\widehat{\eta}_1, \widehat{\eta}_2'\widehat{\eta}_2, \ldots, \widehat{\eta}_r'\widehat{\eta}_r$ are independent under H_0 (compare (4.8.2)). Now Theorem 4.8.1 follows directly.

Theorem 4.8.1. *Under* $H_0 : \{\forall i : C_i\beta_i = 0\}$:

$$(LR_0)^{\frac{2}{N}} \sim \Lambda_{A,D,T,S}, \textit{ with } \begin{array}{ll} a_i = N_i/N_1 & d_i = m_i, \\ t_i = l_{1(i)}, & s_i = r_{(i)}, \quad \textit{for } i = 1, \ldots, r. \end{array} \tag{4.8.4}$$

Denote the double restricted parameterspace of $\theta = (\beta, \Sigma)$ by Θ_{00}. The likelihood ratio LR_{00} for (4.6.3) becomes

$$LR_{00} = \frac{\sup_{\theta\in\Theta_{00}} L(\theta; Y)}{\sup_{\theta\in\Theta_0} L(\theta; Y)|} = \prod_{i=1}^{r} \left(\frac{|\widehat{\eta}_{0i}'\widehat{\eta}_{0i}|}{|\widehat{\eta}_{0i}'\widehat{\eta}_{0i} + \widehat{\nu}_{01i}'\widehat{\nu}_{01i}|} \right)^{\frac{N_i}{2}} = \prod_{i=1}^{r} \Lambda_{00i}^{\frac{N_i}{2}}. \tag{4.8.5}$$

Since $\nu_i' H_{01(i)} \nu_i = 0$ under H_{00}, applying Theorem 4.7.1 to $\widehat{\nu}_{00i} \in L_{00(i)}$, $\widehat{\nu}_{01i} \in L_{01(i)}$ and $\widehat{\eta}_{0i} \in L_{0(i)}^{\perp}$ leads to the conclusion that $\widehat{\nu}_{011}'\widehat{\nu}_{011}, \widehat{\nu}_{012}'\widehat{\nu}_{012}, \ldots, \widehat{\nu}_{01r}'\widehat{\nu}_{01r}$, $\widehat{\eta}_{01}'\widehat{\eta}_{01}, \widehat{\eta}_{02}'\widehat{\eta}_{02}, \ldots, \widehat{\eta}_{0r}'\widehat{\eta}_{0r}$ are independent under H_{00} (compare (4.8.2)). This proves the following generalization of Theorem 4.8.1.

Theorem 4.8.2. *Under* $H_{00} : \{\forall i : C_i\beta_i = 0, D_i\beta_i = 0\}$:

$$(LR_{00})^{\frac{2}{N}} \sim \Lambda_{A,D,T,S}, \quad \textit{with} \quad a_i = N_i/N_1, \quad d_i = m_i, \qquad (4.8.6)$$
$$t_i = l_{01(i)}, \quad s_i = r_{0(i)}, \ \textit{for}\ i = 1, \ldots, r.$$

Note that in both (4.8.4) and (4.8.6) T contains the degrees of freedom of the null hypothesis, while S contains the degrees of freedom of the error terms under the alternative hypothesis.

4.9 A numerical illustration

We now apply the estimation and testing procedures to the numerical example described in Section 4.2. All the tests are performed on a 5% significance level.

The OLS estimation is straightforward by columnwise regression of the dependent variables on only the explanatory variables. To obtain our EGLS estimates, the orthogonal projections described in Section 4.3.4 have to be sequentially performed for groups $i = 1, 2, 3$. For $i = 1$ this gives $\widehat{\mu}_1 = Z_1$, $\widehat{\varepsilon}_1 = E_1$ while $\widehat{\beta}_1$ coincides with the OLS estimate (4.3.5). For $i = 2, 3$, $\widehat{\nu}_i$ follows from (4.3.18), and the EGLS estimates $\widehat{\beta}_i$ and $\widehat{\alpha}_i$ are sequentially determined according to (4.3.20). The EGLS estimate $\widehat{S}$ follows from (4.3.21) and the ML estimate $\widehat{\Sigma}$ is determined according to (4.3.24) and (4.3.26).

We will discuss four tests, of which one in more detail; Table 4.9.1 contains the hypotheses and results for these tests. Assume that we are particularly interested in the testing problem (4.6.1) with $C_i = [0\ 0\ 0\ 1]\forall i$, and in (4.6.3) with $D_i = [0\ 0\ 1\ 0]\forall i$. The estimates for the corresponding restricted and double restricted model are given in Appendix 4.12.2 and 4.12.3. The results for the complete data are presented in Appendices 4.12.6 and 4.12.7 for comparison. Neither the estimation technique nor the missing observations results in large differences in the estimates. The latter phenomenon seems logical in view of the relative small number of missing observations.

Appendix 4.12.4 contains the combined centered MANOVA-tables with the required statistics to perform the two LR tests discussed above. For testing the

significance of the fourth explanatory variable, the LR statistic is determined according to (4.8.3); we found $LR_0^{\frac{2}{N}} = 0.3070$. From the MANOVA-tables and the structure of the dataset, it follows that

$$(LR_0)^{\frac{2}{N}} \sim \Lambda_{[1\ 11/12\ 10/12],[1\ 2\ 1],[1\ 1\ 1],[8\ 6\ 3]}.$$

Since we do not have an analytical expression available yet for the quantiles of the generalized Wilks' distribution, the critical values were determined with simulation (runsize 1,000,000). In Section 4.10 we discuss theoretical approximations for the generalized Wilks' distribution, not based on simulation.

Table 4.9.1 gives the main results for this test (in row 3) and the three other tests. The table contains the null and alternative hypotheses, the values of the corresponding test statistics and the critical values for the performed tests on a 5% significance level. The tests are performed for both the dataset with missing observations and the complete data. In tests 1 through 3, $LR_0^{\frac{2}{N}}$ is the test statistic; in the last test $LR_{00}^{\frac{2}{N}}$.

For the complete data, these test statistics coincide with the usual test statistic Wilks' lambda. (The corresponding critical values are given by *e.g.* Kres (1983), p. 32.) In Table 4.9.1 the abbreviations TS and CV stand for Test Statistic and Critical Value.

Null hypothesis	*Alternative hypothesis*	*Incomplete data*		*Complete data*	
		TS	*CV*	*TS*	*CV*
1. $\forall i : \beta_i = 0$	$\exists i : \beta_i \neq 0$	0.0019	0.0148	0.0018	0.0249
2. $\forall i : \beta_{vi} = 0$	$\exists i : \beta_{vi} \neq 0$	0.0240	0.0262	0.0229	0.0432
3. $\forall i : \beta_{4i} = 0$	$\exists i : \beta_{4i} \neq 0$	0.3070	0.1348	0.3061	0.1940
4. $\forall i : \beta_{3i} = \beta_{4i} = 0$	$\exists i : \beta_{3i} \neq 0\ \forall i : \beta_{4i} = 0$	0.4474	0.2053	0.3156	0.2486

Table 4.9.1: Tests for the numerical example

From the results in Table 4.9.1 it can be concluded that, for example, the null hypothesis 3 of an insignificant fourth explanatory variable is not rejected. The conclusions for all the tests are identical for the complete and incomplete data. This seems (again) logical in view of the relative small number of missing observations.

4.10 Approximating generalized Wilks' distributions

4.10.1 Box transformations

Our approximation for the generalized Wilks' distribution is formulated for the choice $M = r$. This gives no loss of generality because we identify each group to consist of one dependent variable.

In Theorem 4.8.1 we saw that our test statistic $LR_0^{\frac{2}{N}}$ in (4.8.3) has a generalized Wilks' distribution under H_0. In case of complete data, this distribution coincides with the (usual) Wilks' distribution. For the latter, two approximations are well known: the χ^2-distribution of Bartlett (1947) and the F-approximation of Rao (1952). In this section we will approximate the generalized Wilks' distribution by means of χ^2-distributions and compare the different approximations by means of a simulation study.

The approximations can be derived by means of transformations which were introduced in Box (1949); we have used the main result of the transformations as presented in Muirhead (1982) Section 8.2.4. Recall that $l_{(i)}$ denotes the dimension of $L_{(i)} = \mathcal{R}(X_i\, Y_{(i-1)})$, while $a_i = N_i/N$.

Theorem 4.10.1. *Under the null hypothesis H_0 in* (4.6.1)*, a second order approximation of the distribution of $Q = -2log(LR_0^{\frac{2}{N}})$, is*

$$P(Q \leq q) = (1-\omega_2)P(\chi_f^2 \leq \rho q) + \omega_2 P(\chi_{f+4}^2 \leq \rho q) + O(N^{-3}) \quad (4.10.1)$$

with

$$\begin{aligned} f &= \sum_{i=1}^{M} l_{1(i)}, \\ \rho &= \frac{N}{2}\rho_0 = \frac{N}{2}\left[1 - \frac{1}{2Nf}\sum_{i=1}^{r}\frac{l_{1(i)}}{a_i}\left[l_{1(i)} + 2l_{0(i)} + 2\right]\right], \\ \omega_2 &= \frac{1}{12N^2\rho_0^2}\sum_{i=1}^{r}\frac{l_{1(i)}}{a_i^2}\left[3l_{0(i)}(2 + l_{(i)}) + (l_{1(i)} + 2)(l_{1(i)} + 1)\right] - \frac{(1-\rho_0)^2}{4\rho_0^2}f. \end{aligned}$$

Proof. Since $M = r$ we have $m_i = 1$ and so $\Lambda_{0i} \sim Beta(\frac{1}{2}r_{(i)}, \frac{1}{2}l_{0(i)})$. The moments of LR_0 follow from its definition (4.8.3) and from the independence and

the moments of Λ_{0i}:

$$E\{LR_0^h\} = E\{\prod_{i=1}^{r} \Lambda_{0i}^{\frac{N_i h}{2}}\} = K \prod_{i=1}^{r} \frac{\Gamma\left[\frac{1}{2}N_i(1+h) - \frac{1}{2}l_{(i)}\right]}{\Gamma\left[\frac{1}{2}N_i(1+h) - \frac{1}{2}l_{0(i)}\right]}, \tag{4.10.2}$$

where K is a constant not involving h. Box transformations applied to LR_0 lead, after algebraic manipulations, to the approximating distribution (4.10.1) with parameters f, ρ_0 and ω_2 (see Appendix 4.12.8). Since $log(LR_0^{\frac{2}{N}}) = \frac{2}{N} log(LR_0)$, the approximating distribution of the logarithm of the test statistic is identical to the one of the logarithm of the likelihood ratio, except the scale parameter ρ. The scale parameter of the test statistic is $\frac{N}{2}$ times ρ_0. □

In case of only the constant as explanatory variable ($l_{1(i)} = 1$ and $l_{0(i)} = i-1$), our parameters reduce to the ones derived in Bhargava (1962). We call (4.10.1) the Box approximation.

An approximation of the distribution of the test statistic $LR_{00}^{\frac{2}{N}}$ can be derived in a similar way.

Corollary. *Under H_{00} in* (4.6.3)*, the second order approximation of the distribution of $Q = -2log(LR_{00}^{\frac{2}{N}})$ is equal to (4.10.1) with the parameters $l_{(i)}$ and $l_{1(i)}$ replaced by $l_{0(i)}$ and $l_{01(i)}$, respectively.*

From (4.10.1) the first order approximation follows

$$P(Q \leq q) = P(\chi_f^2 \leq \rho q) + O(N^{-2}). \tag{4.10.3}$$

Since (4.10.3) coincides with Bartlett's approximation in case of complete data, we will call (4.10.3) Bartlett's approximation even in this more general situation.

4.10.2 A simulation study

We compare approximations (4.10.1), (4.10.3) and the standard approximation (*i.e.* $-2log(LR_0) \sim \chi_f^2$) by means of simulation (with runsize 1,000,000). First the critical value of our test statistic (with significance level α) is determined by means of simulation. Then the probability that our test statistic exceeds this critical value is determined according to the three different approximations. This has

been done under the assumption that there are four explanatory variables, three groups and p linear constraints per group ($p_i = p$ for all i). The simulations have been performed for different values of the significance level α, number of cases (N), number of constraints p, fractions of missing data ($A = [a_1\ a_2\ a_3]$ with $a_i = N_i/N$) and different number of variables per group ($D = [m_1\ m_2\ m_3]$). Table 4.10.1 contains the results for $D = [1\ 2\ 1]$.

$D = [1 \quad 2 \quad 1]$		$A = [1 \quad 0.9 \quad 0.8]$			$A = [1 \quad 0.8 \quad 0.6]$		
$\alpha = 0.05$		Standard	Bartlett	Box	Standard	Bartlett	Box
$N = 20$	$p = 1$	.009	.047	.050	.004	.040	.048
	$p = 2$	.012	.047	.050	.007	.042	.049
	$p = 4$	.037	.047	.050	.032	.045	.049
$N = 200$	$p = 1$	.044	.050	.050	.044	.050	.050
	$p = 2$	.045	.050	.050	.044	.050	.050
	$p = 4$	.049	.050	.050	.048	.050	.050
$N = 2000$	$p = 1$	.049	.050	.050	.049	.050	.050
	$p = 2$	.049	.050	.050	.049	.050	.050
	$p = 4$	.050	.050	.050	.050	.050	.050
$\alpha = 0.10$							
$N = 20$	$p = 1$	.026	.096	.100	.015	.085	.098
	$p = 2$	.031	.095	.100	.020	.087	.098
	$p = 4$	.078	.095	.100	.070	.092	.099
$N = 200$	$p = 1$	.091	.100	.100	.090	.100	.100
	$p = 2$	.092	.100	.100	.090	.100	.100
	$p = 4$	.098	.100	.100	.097	.100	.100
$N = 2000$	$p = 1$	.099	.100	.100	.099	.100	.100
	$p = 2$	.099	.100	.100	.099	.100	.100
	$p = 4$	.100	.100	.100	.100	.100	.100

Table 4.10.1: Simulated approximations for $D = [1 \quad 2 \quad 1]$

As can be expected, the accuracy of the approximations increases with the sample sizes. Approximation (4.10.1) outperforms the other ones. The standard approximation is quite bad for small sample sizes. Only for $N = 2000$, this

approximation gives good results. Approximation (4.10.3) performs well for big sample sizes ($N = 200(0)$), but is not as accurate as approximation (4.10.1) for small sample sizes ($N = 20$). All the approximations seem to improve with the number of constraints (p). As the fraction of missing observations increases, the approximations become less accurate.

To study the effect of the number of variables per group on the quality of the approximations, we also did a simulation for $D = [1\ 3\ 2]$. Table 4.10.2 contains the results.

$D = [1 \quad 3 \quad 2]$		$A = [1 \quad 0.9 \quad 0.8]$			$A = [1 \quad 0.8 \quad 0.6]$		
$\alpha = 0.05$		Standard	Bartlett	Box	Standard	Bartlett	Box
$N = 20$	$p = 1$	.003	.040	.049	.000	.022	.040
	$p = 2$	.003	.041	.049	.001	.027	.043
	$p = 4$	.017	.042	.049	.001	.035	.046
$N = 200$	$p = 1$	.042	.050	.050	.040	.050	.050
	$p = 2$	.046	.050	.050	.041	.050	.050
	$p = 4$	.049	.050	.050	.045	.050	.050
$N = 2000$	$p = 1$	.049	.050	.050	.049	.050	.050
	$p = 2$	.049	.050	.050	.049	.050	.050
	$p = 4$	.049	.050	.050	.050	.050	.050
$\alpha = 0.10$							
$N = 20$	$p = 1$	.009	.085	.098	.002	.054	.086
	$p = 2$	.011	.086	.099	.003	.063	.091
	$p = 4$	.041	.087	.099	.026	.077	.096
$N = 200$	$p = 1$	.088	.100	.100	.085	.100	.100
	$p = 2$	.087	.100	.100	.085	.100	.100
	$p = 4$	.094	.100	.100	.092	.100	.100
$N = 2000$	$p = 1$	.098	.100	.100	.098	.100	.100
	$p = 2$	.098	.100	.100	.099	.100	.100
	$p = 4$	.100	.100	.100	.099	.100	.100

Table 4.10.2: Simulated approximations for $D = [1 \quad 3 \quad 2]$

The previous conclusions about the effect of the different parameters still re-

main valid. However, in comparison to Table 4.10.1, the quality of the approximations is worse if there is only a small number of observations ($N = 20$) available.

4.11 Conclusions and further research

This chapter discussed estimation and testing for a linear regression model with complete observations for the explanatory variables and consecutively added dependent variables, leading to a specific incomplete data structure. For this model, OLS and GLS do not longer coincide, so we discussed EGLS. A specific choice of EGLS estimation, which coincides with ML estimation, was analysed in detail. Exact tests for restricted and double restricted models were presented. Different approximations of the distribution of the test statistic were compared.

The relative efficiency of the OLS estimators in relation to the (E)GLS estimators for the regression coefficients have been discussed in more detail. The small sample properties of the remaining estimators have not been analysed in detail yet. Especially the first step of EGLS estimation, *i.e.* the choice of the covariance estimator, is interesting for further research.

The LR test for linear restrictions on the regression coefficients under the normality assumptions has been extensively discussed. Other well known test statistics for complete data, are the test statistics of Pillai, Hotelling and Roy. The derivation of similar test statistics for incomplete data is left for further research. It could also be interesting to look at a similar test as the one which was constructed by Krishnamoorthy and Pannala (1998) for the model with only the constant term as explanatory variable.

4.12 Appendices

4.12.1 Missing data: the unrestricted model

OLS estimates

$$b = \begin{bmatrix} 2.0000 & 5.0000 & 5.0000 & 3.4107 \\ 1.0000 & -1.0000 & 1.0000 & 0.9821 \\ 1.0000 & 2.0000 & 0.0000 & 0.1964 \\ -1.0000 & -1.0000 & -1.0000 & -1.0536 \end{bmatrix}$$

$$S = \begin{bmatrix} 2.2500 & 1.2027 & 2.4054 & -0.6959 \\ 1.2027 & 2.5714 & 0.0000 & -0.0496 \\ 2.4045 & 0.0000 & 10.2857 & -2.7775 \\ -0.6959 & -0.0496 & -2.7775 & 2.1964 \end{bmatrix}$$

EGLS estimates

$$\widehat{\beta} = \begin{bmatrix} 2.0000 & 5.4091 & 5.8182 & 3.1919 \\ 1.0000 & -1.0000 & 1.0000 & 0.9815 \\ 1.0000 & 1.8636 & -0.2727 & 0.2694 \\ -1.0000 & -0.9545 & -0.9091 & -1.0774 \end{bmatrix}$$

$$\widehat{S} = \begin{bmatrix} 2.2500 & 1.2756 & 2.5511 & -0.7382 \\ 1.2756 & 2.6246 & 0.1063 & -0.0951 \\ 2.5511 & 0.1063 & 10.4982 & -2.8377 \\ -0.7382 & -0.0951 & -2.8377 & 2.2139 \end{bmatrix}$$

ML estimate

$$\widehat{\Sigma} = \begin{bmatrix} 1.5000 & 1.0227 & 2.0455 & -0.5480 \\ 1.0227 & 1.7758 & 0.2789 & -0.1050 \\ 2.0455 & 0.2789 & 7.1033 & -1.7858 \\ -0.5480 & -0.1050 & -1.7858 & 1.3169 \end{bmatrix}$$

4.12.2 Missing data: the restricted model

OLS estimates

$$b_0 = \begin{bmatrix} 4.0000 & 7.0769 & 7.0769 & 5.4839 \\ -0.3333 & -2.3187 & -0.3187 & -0.4113 \\ 0.6667 & 1.6374 & -0.3626 & -0.1774 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

$$S_0 = \begin{bmatrix} 3.3333 & 2.5526 & 3.6132 & 0.9083 \\ 2.5526 & 3.7335 & 1.4835 & 1.1833 \\ 3.6132 & 1.4835 & 10.4835 & -0.8794 \\ 0.9083 & 1.1833 & -0.8794 & 3.6014 \end{bmatrix}$$

EGLS estimates

$$\widehat{\beta}_0 = \begin{bmatrix} 4.0000 & 7.3889 & 7.5185 & 5.6474 \\ -0.3333 & -2.2593 & -0.2346 & -0.3881 \\ 0.6667 & 1.5185 & -0.5309 & -0.2238 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

$$\widehat{S}_0 = \begin{bmatrix} 3.3333 & 2.6022 & 3.6835 & 0.9496 \\ 2.6022 & 3.7762 & 1.5439 & 1.2442 \\ 3.6835 & 1.5439 & 10.5690 & -0.8337 \\ 0.9496 & 1.2442 & -0.8337 & 3.6166 \end{bmatrix}$$

ML estimate

$$\widehat{\Sigma}_0 = \begin{bmatrix} 2.5000 & 2.0278 & 2.8704 & 0.7295 \\ 2.0278 & 2.7629 & 1.1464 & 0.9570 \\ 2.8704 & 1.1464 & 7.7199 & -0.5300 \\ 0.7295 & 0.9570 & -0.5300 & 2.4869 \end{bmatrix}$$

4.12.3 Missing data: the double restricted model

OLS estimates

$$b_{00} = \begin{bmatrix} 5.0000 & 9.6818 & 6.5000 & 5.2000 \\ 0.0000 & -1.5000 & -0.5000 & -0.5000 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

$$S_{00} = \begin{bmatrix} 3.6000 & 4.0247 & 2.8460 & 0.6037 \\ 4.0247 & 7.0152 & 0.5000 & 0.6128 \\ 2.8460 & 0.5000 & 9.5000 & -0.6835 \\ 0.6037 & 0.6128 & -0.6835 & 3.2000 \end{bmatrix}$$

EGLS estimates

$$\widehat{\beta}_{00} = \begin{bmatrix} 5.0000 & 9.7813 & 6.5703 & 5.1376 \\ 0.0000 & -1.5000 & -0.5000 & -0.5000 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

$$\widehat{S}_{00} = \begin{bmatrix} 3.6000 & 4.0352 & 2.8535 & 0.5898 \\ 4.0352 & 7.0272 & 0.5085 & 0.5848 \\ 2.8535 & 0.5085 & 9.5060 & -0.6961 \\ 0.5898 & 0.5848 & -0.6961 & 3.2049 \end{bmatrix}$$

ML estimate

$$\widehat{\Sigma}_{00} = \begin{bmatrix} 3.0000 & 3.2812 & 2.3203 & 0.3876 \\ 3.2812 & 5.5320 & 0.2623 & 0.2392 \\ 2.3203 & 0.2623 & 7.6689 & -0.6188 \\ 0.3876 & 0.2392 & -0.6188 & 2.5704 \end{bmatrix}$$

4.12.4 The collection of centered MANOVA-tables

Group 1		
Space	**SS**	**DF**
$\widetilde{L}_{00(1)}$	0	1
$L_{01(1)}$	6	1
$\widetilde{L}_{0(1)}$	6	2
$L_{1(1)}$	12	1
$\widetilde{L}_{(1)}$	18	3
$L^{\perp}_{(1)}$	18	8
$\mathcal{R}(1_{12})^{\perp}$	36	11
$\mathcal{R}(1_{12})$	300	1
$I\!R^{12}$	336	12

Group 2		
Space	**SS**	**DF**
$\widetilde{L}_{00(2)}$	$\begin{bmatrix} 82.2614 & 43.0313 \\ 43.0313 & 25.3828 \end{bmatrix}$	2
$L_{01(2)}$	$\begin{bmatrix} 9.0750 & -12.0313 \\ -12.0313 & 15.9505 \end{bmatrix}$	1
$\widetilde{L}_{0(2)}$	$\begin{bmatrix} 91.3364 & 31.0000 \\ 31.0000 & 41.3333 \end{bmatrix}$	3
$L_{1(2)}$	$\begin{bmatrix} 0.4364 & -0.7273 \\ -0.7273 & 1.2121 \end{bmatrix}$	1
$\widetilde{L}_{(2)}$	$\begin{bmatrix} 91.7727 & 30.2727 \\ 30.2727 & 42.5455 \end{bmatrix}$	4
$L^{\perp}_{(2)}$	$\begin{bmatrix} 11.8636 & -12.2727 \\ -12.2727 & 47.4545 \end{bmatrix}$	6
$\mathcal{R}(1_{11})^{\perp}$	$\begin{bmatrix} 103.6364 & 18.0000 \\ 18.0000 & 90.0000 \end{bmatrix}$	10
$\mathcal{R}(1_{11})$	$\begin{bmatrix} 295.3636 & 285.0000 \\ 285.0000 & 275.0000 \end{bmatrix}$	1
$I\!R^{11}$	$\begin{bmatrix} 399 & 303 \\ 303 & 365 \end{bmatrix}$	11

Group 3		
Space	**SS**	**DF**
$\widetilde{L}_{00(3)}$	9.0849	4
$L_{01(3)}$	2.5022	1
$\widetilde{L}_{0(3)}$	11.5872	5
$L_{1(3)}$	9.8462	1
$\widetilde{L}_{(3)}$	21.4333	6
$L^{\perp}_{(3)}$	8.6667	3
$\mathcal{R}(1_{10})^{\perp}$	30.1000	9
$\mathcal{R}(1_{10})$	136.9000	1
$I\!R^{10}$	167	10

$$LR_0 = 0.3070, \quad LR_{00} = 0.4474.$$

4.12.5 Complete data: the unrestricted model

LS estimates

$$b = \begin{bmatrix} 2.0000 & 5.0000 & 5.0000 & 4.0000 \\ 1.0000 & -1.0000 & 1.0000 & 1.0000 \\ 1.0000 & 2.0000 & 0.0000 & 0.0000 \\ -1.0000 & -1.0000 & -1.0000 & -1.0000 \end{bmatrix}$$

$$S = \begin{bmatrix} 2.2500 & 1.1250 & 2.2500 & 0.0000 \\ 1.1250 & 2.2500 & 0.0000 & 0.0000 \\ 2.2500 & 0.0000 & 9.0000 & -2.2500 \\ 0.0000 & 0.0000 & -2.2500 & 2.2500 \end{bmatrix}$$

ML estimate

$$\widehat{\Sigma} = \begin{bmatrix} 1.5000 & 0.7500 & 1.5000 & 0.0000 \\ 0.7500 & 1.5000 & 0.0000 & 0.0000 \\ 1.5000 & 0.0000 & 6.0000 & -1.5000 \\ 0.0000 & 0.0000 & -1.5000 & 1.5000 \end{bmatrix}$$

4.12.6 Complete data: the restricted model

LS estimates

$$b_0 = \begin{bmatrix} 4.0000 & 7.0000 & 7.0000 & 6.0000 \\ -0.3333 & -2.3333 & -0.3333 & -0.3333 \\ 0.6667 & 1.6667 & -0.3333 & -0.3333 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

$$S_0 = \begin{bmatrix} 3.3333 & 2.3333 & 3.3333 & 1.3333 \\ 2.3333 & 3.3333 & 1.3333 & 1.3333 \\ 3.3333 & 1.3333 & 9.3333 & -0.6667 \\ 1.3333 & 1.3333 & -0.6667 & 3.3333 \end{bmatrix}$$

ML estimate

$$\widehat{\Sigma}_0 = \begin{bmatrix} 2.5000 & 1.7500 & 2.5000 & 1.0000 \\ 1.7500 & 2.5000 & 1.0000 & 1.0000 \\ 2.5000 & 1.0000 & 7.0000 & -0.5000 \\ 1.0000 & 1.0000 & -0.5000 & 2.5000 \end{bmatrix}$$

4.12.7 Complete data: the double restricted model

LS estimates

$$b_{00} = \begin{bmatrix} 5.0000 & 9.5000 & 6.5000 & 5.5000 \\ 0.0000 & -1.5000 & -0.5000 & -0.5000 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

$$S_{00} = \begin{bmatrix} 3.6000 & 3.6000 & 2.7000 & 0.9000 \\ 3.6000 & 6.7500 & 0.4500 & 0.4500 \\ 2.7000 & 0.4500 & 8.5500 & -0.4500 \\ 0.9000 & 0.4500 & -0.4500 & 3.1500 \end{bmatrix}$$

ML estimate

$$\widehat{\Sigma}_{00} = \begin{bmatrix} 3.0000 & 3.0000 & 2.2500 & 0.7500 \\ 3.0000 & 5.6250 & 0.3750 & 0.3750 \\ 2.2500 & 0.3750 & 7.1250 & -0.3750 \\ 0.7500 & 0.3750 & -0.3750 & 2.6250 \end{bmatrix}$$

4.12.8 Box transformations

To approximate the generalized Wilks' distribution, we have used the main result of Box transformations as presented in Muirhead (1982) Section 8.2.4:

Consider a random variable Z ($0 \leq Z \leq 1$) with moments:

$$E\{Z^h\} = K \left[\frac{\prod_{j=1}^{p} y_j^{y_j}}{\prod_{k=1}^{q} x_k^{x_k}} \right]^h \frac{\prod_{k=1}^{q} \Gamma\left[x_k(1+h)+\xi_k\right]}{\prod_{j=1}^{p} \Gamma\left[y_j(1+h)+\eta_j\right]},$$

where

$$\sum_{j=1}^{p} y_j = \sum_{k=1}^{q} x_k$$

and K is a constant such that E$\{Z^0\}$=1. Then

$$P(-2\rho log(Z) \leq x) =$$

$$P(\chi_f^2 \leq x) + \omega_2 \left[P(\chi_{f+4}^2 \leq x) - P(\chi_f^2 \leq x)\right] + O(N^{-3}),$$

where

$$f = -2\left[\sum_{k=1}^{q}\xi_k - \sum_{j=1}^{p}\eta_j - \frac{1}{2}(q-p)\right]$$

and

$$\rho = 1 - \frac{1}{f}\left[\sum_{k=1}^{q} x_k^{-1}\left(\xi_k^2 - \xi_k + \frac{1}{6}\right) - \sum_{j=1}^{p} y_j^{-1}\left(\eta_j^2 - \eta_j + \frac{1}{6}\right)\right]$$

and

$$\begin{aligned}\omega_2 &= -\frac{1}{6\rho^2}\left\{\sum_{k=1}^{q} x_k^{-2}\left[(\beta_k+\xi_k)^3 - \frac{3}{2}(\beta_k+\xi_k)^2 + \frac{1}{2}(\beta_k+\xi_k)\right]\right. \\ &\quad \left. - \sum_{j=1}^{p} y_j^{-2}\left[(\epsilon_j+\eta_j)^3 - \frac{3}{2}(\epsilon_j+\eta_j)^2 + \frac{1}{2}(\epsilon_j+\eta_j)\right]\right\},\end{aligned}$$

with

$$\beta_k = (1-\rho)x_k, \quad \epsilon_j = (1-\rho)y_j.$$

Since the moments of our test statistic LR_0 have that specific shape (see (4.10.2)), Box transformations can be applied and give

$$f = -2\left[\sum_{i=1}^{r}(-\frac{1}{2}l_{(i)} + \frac{1}{2}l_{0(i)})\right] = \sum_{i=1}^{r} l_{1(i)}$$

and

$$\begin{aligned}\rho_0 &= 1 - \frac{1}{f}\sum_{i=1}^{r}\frac{2}{N_i}\left[\left\{(-\frac{1}{2}l_{(i)})^2 + \frac{1}{2}l_{(i)} + \frac{1}{6}\right\} - \left\{(-\frac{1}{2}l_{0(i)})^2 + \frac{1}{2}l_{0(i)} + \frac{1}{6}\right\}\right] \\ &= 1 - \frac{1}{f}\sum_{i=1}^{r}\frac{1}{N_i}\left[\left\{\frac{1}{2}(l_{0(i)}^2 + l_{1(i)}^2 + 2l_{0(i)}l_{1(i)}) - \frac{1}{2}l_{0(i)}^2 + \frac{1}{2}l_{1(i)}\right\}\right] \\ &= 1 - \frac{1}{2f}\sum_{i=1}^{r}\frac{l_{1(i)}}{N_i}\left[l_{1(i)} + 2l_{0(i)} + 2\right]\end{aligned}$$

and

$$\begin{aligned}
\omega_2 &= -\frac{1}{6\rho_0^2}\sum_{i=1}^{r}\frac{4}{N_i^2}\left[\left\{(\epsilon_i+\gamma_i-\frac{1}{2}l_{1(i)})^3-\frac{3}{2}(\epsilon_i+\gamma_i-\frac{1}{2}l_{1(i)})^2+\right.\right.\\
&\qquad\left.\left.\frac{1}{2}(\epsilon_i+\gamma_i-\frac{1}{2}l_{1(i)})-(\epsilon_i+\gamma_i)^3-\frac{3}{2}(\epsilon_i+\gamma_i)^2+\frac{1}{2}(\epsilon_i+\gamma_i)\right\}\right]\\
&= -\frac{1}{6\rho_0^2}\sum_{i=1}^{r}\frac{l_{1(i)}}{N_i^2}\left[-6(\epsilon_i+\gamma_i)^2+3(\epsilon_i+\gamma_i)l_{1(i)}+6(\epsilon_i+\gamma_i)-\right.\\
&\qquad\left.\frac{1}{2}l_{1(i)}^2-\frac{3}{2}l_{1(i)}-1\right]\\
&= -\frac{1}{6\rho_0^2}\sum_{i=1}^{r}\frac{l_{1(i)}}{N_i^2}\left[3(\epsilon_i+\gamma_i)(-2(\epsilon_i+\gamma_i)+l_{1(i)}+2)-\frac{1}{2}(l_{1(i)}^2+3l_{1(i)}+2)\right]\\
&= -\frac{1}{12\rho_0^2}\sum_{i=1}^{r}\frac{l_{1(i)}}{N_i^2}\left[3((1-\rho_0)N_i-l_{0(i)})(2+l_{(i)}-(1-\rho_0)N_i)-\right.\\
&\qquad\left.(l_{1(i)}+2)(l_{1(i)}+1)\right]\\
&= -\frac{1}{12\rho_0^2}\left[-3(1-\rho_0)^2\sum_{i=1}^{r}l_{1(i)}+3(1-\rho_0)\sum_{i=1}^{r}\frac{l_{1(i)}}{N_i}\left[l_{1(i)}+2l_{0(i)}+2\right]\right.\\
&\qquad\left.-\sum_{i=1}^{r}\frac{l_{1(i)}}{N_i^2}\left[3l_{0(i)}(2+l_{(i)})+(l_{1(i)}+2)(l_{1(i)}+1)\right]\right]\\
&= -\frac{(1-\rho_0)^2}{4\rho_0^2}f_0+\frac{1}{12\rho_0^2}\sum_{i=1}^{r}\frac{l_{1(i)}}{N_i^2}\left[3l_{0(i)}(2+l_{(i)})+(l_{1(i)}+2)(l_{1(i)}+1)\right].
\end{aligned}$$

Chapter 5

Additional topics of multivariate regression

5.1 Introduction

The previous chapter introduced the model for multivariate regression with consecutively added dependent variables. Several estimators were presented and the distribution of the test statistic (based on the likelihood ratio) was derived. Some additional features of this model will be discussed in this chapter.

In Section 5.2 we will introduce new classes of covariance estimators and prove consistency of these estimators and of the estimators presented in Chapter 4.

We further discuss two, widely used, alternative estimation techniques for our model: iterative EGLS (in Section 5.3) and the EM-algorithm (in Section 5.4). Unlike the estimation procedure of Chapter 4, these iterative procedures do not result in closed form estimators for the coefficients.

In Section 4.3.4 it was shown that for EGLS estimation the dependent variables are used in a well-structured way. For the model with the constant term as the sole explanatory variable, this resulted in nice expressions for the EGLS estimators (see Section 4.5.2). In Section 5.5 we look at a simple generalization: one-way MANOVA. For this model, the usual MANOVA-tables (for complete data) must be adapted in a non-trivial way.

The final Section 5.6 reviews and discusses our results.

5.2 Consistency of estimators

5.2.1 Introduction

We consider the asymptotic behavior for $N_r \to \infty$. Since $N_i \geq N_{i+1}$, this implies $N_i \to \infty$ for all i. Without loss of generality, we take $m_i = 1$ for all i throughout this section.

In the notation of random variables Z_{N_r} depending on N_r, we omit the subindex N_r for greater readability. As usual the notation $Z = O_P(N_r)$ for a random vector Z means that Z is of order N_r in probability:

$$\sup_{N_r} P(N_r^{-1}|Z| \geq z) \to 0 \text{ as } z \to \infty.$$

To enhance the readability of the proofs, we will sometimes use the additional notation $Z = o_P(N_r)$ for $\frac{1}{N_r} Z \xrightarrow{P} 0$.

We make the following three assumptions

$$N_1 = O(N_r), \tag{5.2.1}$$

$$\text{the } \varepsilon_{t(r)} \text{ are i.i.d.}, \tag{5.2.2}$$

$$(X_r' X_r)^{-1} \to 0. \tag{5.2.3}$$

The first assumption implies that $O(N_r)$, $O_P(N_r)$ and $o_P(N_r)$ are equivalent to $O(N_i)$, $O_P(N_i)$ and $o_P(N_i)$, respectively. So all samples sizes increase in more or less the same way to infinity.

As a consequence of (5.2.1) and (5.2.2), the law of large numbers can be applied to all groups i:

$$\frac{1}{N_h} \sum_{t=1}^{N_h} \varepsilon_{t(i)} \varepsilon_{t(i)}' \xrightarrow{P} \Sigma_{(i)(i)}, \quad \text{for } h = i, \ldots, r, \tag{5.2.4}$$

$$\frac{1}{N_h} \sum_{t=1}^{N_h} \eta_{ti} \eta_{ti}' \xrightarrow{P} \Gamma_{ii}, \quad \text{for } h = i, \ldots, r, \tag{5.2.5}$$

$$\frac{1}{N_h} \sum_{t=1}^{N_h} \zeta_{ti} \zeta_{ti}' \xrightarrow{P} \Delta_{ii}, \quad \text{for } h = i, \ldots, r. \tag{5.2.6}$$

We only prove consistency of the unrestricted estimators (of the previous chapter), since the proofs for the restricted estimators are quite similar. In the proofs we will extensively use the following properties:

Lemma 5.2.1. *For* $i = 1, \ldots, r$

$$|E_i - \varepsilon_i| = O_P(1), \tag{5.2.7}$$

$$|\varepsilon_i|,\ |\zeta_i| \text{ and } |\eta_i| \text{ are } O_P(N_r^{\frac{1}{2}}). \tag{5.2.8}$$

Proof. Since $E\{|H_i\varepsilon_i|^2\} = tr(E\{\varepsilon_i' H_i \varepsilon_i\}) = tr(H_i E\{\varepsilon_i \varepsilon_i'\}) = l_i\sigma_{ii} \leq k\sigma_{ii}$ and $\varepsilon_i - E_i = H_i\varepsilon_i$, we have $|E_i - \varepsilon_i| = O_P(1)$. Since $E\{|\varepsilon_i|^2\} = N_i\sigma_{ii}$, we have $|\varepsilon_i| = O_P(N_r^{\frac{1}{2}})$. Similarly, $E\{|\zeta_i|^2\} = N_i\Delta_{ii}$ and $E\{|\eta_i|^2\} = N_i\Gamma_{ii}$ (since $m_i = 1$). □

Omitting a finite or even infinite number of vector elements, while still keeping an infinite number, does not invalidate the lemma. More precisely, let us define

$a^{(h)}$: the first N_h elements of the vector a.

Since

$$|a^{(h)}| \leq |a|, \tag{5.2.9}$$

the following Lemma results directly from (5.2.1) and Lemma 5.2.1.

Lemma 5.2.2. *For* $i = 1, \ldots, r$

$$|E_i^{(h)} - \varepsilon_i^{(h)}| = O_P(1), \tag{5.2.10}$$

$$|\varepsilon_i^{(h)}|,\ |\zeta_i^{(h)}| \text{ and } |\eta_i^{(h)}| \text{ are } O_P(N_r^{\frac{1}{2}}). \tag{5.2.11}$$

5.2.2 OLS

In discussing the consistency of estimators for the regression coefficients, we assume that $r(X_r) = k$. As a consequence $r(X_i) = k$ for all i. For the consistency of the covariance estimators, this assumption is not necessary.

We will denote the matrix of all OLS estimators b_i in (4.3.5) by b. A more precise notation would be b_{N_r} but we drop the subindex (see Section 5.2.1).

Theorem 5.2.3. $b \xrightarrow{P} \beta$.

Proof. From (4.3.5) it follows that

$$b_i - \beta_i = (X_i'X_i)^{-1}X_i'(X_i\beta_i + \varepsilon_i) - \beta_i = G_iX_i'\varepsilon_i.$$

We have $G_i = (X_i'X_i)^{-1} \leq (X_r'X_r)^{-1} \to 0$ by (5.2.3). Therefore

$$E\{(b_i - \beta_i)(b_i - \beta_i)'\} = \sigma_{ii}G_i \to 0,$$

which completes the proof. □

In discussing the consistency of covariance estimators for Σ, we do not assume that $r(X_r) = k$. We look at a broad class of estimators based on the OLS residuals. In this class, the estimators for Σ have entries

$$S_{ig}^{(h)} = \frac{E_i^{(h)'}E_g^{(h)}}{N_h}, \quad g = 1, \ldots, i, \quad i = 1, \ldots, r, \quad \text{with } h \in \{i, \ldots, r\}.$$

The covariance estimators differ in the number of residuals on which they are based: $S_{ig}^{(h)}$ is based on the first N_h OLS residuals of dependent variables i and g.

In practice there are two often used estimators in this class. One of these uses all available residuals ($S_{ig}^{(i)}$ for all i), the other uses only the first N_r residuals ($S_{ig}^{(r)}$ for all i) and discards all the residuals of incomplete observations. These estimators differ in efficiency and positive definiteness but the next theorem states that both are consistent.

Theorem 5.2.4. $S_{ig}^{(h)} \xrightarrow{P} \sigma_{ig}$.

Proof. We have

$$\begin{aligned} |E_i^{(h)'}E_g^{(h)} - \varepsilon_i^{(h)'}\varepsilon_g^{(h)}| &= |E_i^{(h)'}(E_g^{(h)} - \varepsilon_g^{(h)}) + (E_i^{(h)} - \varepsilon_i)^{(h)'}\varepsilon_g^{(h)}| \\ &\leq |E_i^{(h)}||E_g^{(h)} - \varepsilon_g^{(h)}| + |E_i^{(h)} - \varepsilon_i^{(h)}||\varepsilon_g^{(h)}| = O_P(N_r^{\frac{1}{2}}), \end{aligned}$$

where the last equality follows from (5.2.10) and (5.2.11). According to (5.2.4), we have $\frac{1}{N_h}\varepsilon_i^{(h)'}\varepsilon_g^{(h)} \xrightarrow{P} \sigma_{ig}$. Together this implies $\frac{1}{N_h}E_i^{(h)'}E_g^{(h)} \xrightarrow{P} \sigma_{ig}$. □

Corollary. $S \xrightarrow{P} \Sigma$.

This corollary for S in (4.3.7) follows directly from Theorem 5.2.4, since $S_{ig} = \frac{N_i}{r_i}S_{ig}^{(i)}$ and $\frac{N_i}{r_i} \to 1$.

5.2.3 GLS

In this (and next) subsection we present direct proofs of the consistency of (E)GLS estimators instead of verifying the general regularity conditions for consistency of (E)GLS (see Mittelhammer *et al.* (1996) p. 347 and p. 374 *e.g.*). We assume non-collinearity (see Section 5.2.2). Denoting the matrix of all GLS estimators $\breve{\beta}_i$ in (4.3.16) by $\breve{\beta}$, we will show that $\breve{\beta}$ is consistent for β.

Theorem 5.2.5. $\breve{\beta} \xrightarrow{P} \beta$.

Proof. We prove this theorem by using an induction argument. For $i = 1$, GLS and OLS estimation coincide. So according to Theorem 5.2.3 $\breve{\beta}_1 \xrightarrow{P} \beta_1$. For general $i\ (= 2, \ldots, r)$, the induction assumption is $\breve{\beta}_{(i-1)} \xrightarrow{P} \beta_{(i-1)}$. We have

$$\begin{aligned} \breve{\beta}_i - G_i X_i'(Y_i - \zeta_i) &= G_i X_i'(\zeta_i - \breve{\zeta}_i) = G_i X_i'(\breve{\mu}_{(i-1)} - \mu_{(i-1)})\alpha_i \\ &= (\breve{\beta}_{(i-1)} - \beta_{(i-1)})\alpha_i \xrightarrow{P} 0. \end{aligned}$$

The first equality follows from (4.3.16), the second from (4.3.1) and (4.3.12). The convergence in probability follows from the induction assumption.

Furthermore, relations (4.3.1) and (4.3.12) give

$$G_i X_i'(Y_i - \zeta_i) - \beta_i = G_i X_i'(\varepsilon_i - \zeta_i) = G_i X_i' \eta_i \xrightarrow{P} 0$$

since $E\{|G_i X_i' \eta_i|^2\} = \Gamma_{ii} tr\,(G_i) \to 0$. Together, the consistency property

$$\breve{\beta}_i - \beta_i = \Big(\breve{\beta}_i - G_i X_i'(Y_i - \zeta_i)\Big) + (G_i X_i'(Y_i - \zeta_i) - \beta_i) \xrightarrow{P} 0$$

follows. □

We will use this theorem in proving consistency of the EGLS estimators for the regression coefficients.

5.2.4 EGLS

For EGLS we have to minimize (4.2.5) where the covariance-matrix Σ is replaced by a starting estimator, usually obtained with OLS. Different starting estimators

will in general lead to different EGLS estimators for β. In the previous chapter we looked at two specific kinds of EGLS: OLS (in Section 4.3.2) and ML (in Section 4.3.5). Here we will consider general EGLS estimators obtained from a starting estimator $\tilde{S}_0$; they will be denoted by replacing the (ML) superscript $\hat{}$ by $\tilde{}$, like $\tilde{\beta}$ and $\tilde{S}$.

The starting estimator $\tilde{S}_0$ for Σ influences the EGLS estimators only through the resulting starting estimators $\tilde{\alpha}_{0i}$ for α_i; the α_i are specific functions of Σ, see relation (4.3.8), and the $\tilde{\alpha}_{0i}$ are the corresponding functions of $\tilde{S}_0$. (Note that in this section, the subindex 0 indicates the starting estimator and *not* estimators under linear constrictions as in Section 4.6.)

The EGLS estimators for the regression coefficients are very similar to the GLS estimators (4.3.16):

$$\tilde{\beta}_i = G_i X_i'(Y_i - \tilde{\zeta}_i) = G_i X_i'(Y_i - \tilde{\varepsilon}_{(i-1)}\tilde{\alpha}_{0i}). \tag{5.2.12}$$

The EGLS estimator $\tilde{\beta} = [\tilde{\beta}_1 \dots \tilde{\beta}_r]$ turns out to be consistent if the $\tilde{\alpha}_{0i}$ are.

Theorem 5.2.6. *If $\tilde{\alpha}_{0i} \xrightarrow{P} \alpha_i$ for $i = 1, \dots, r$, then $\tilde{\beta} \xrightarrow{P} \beta$.*

Proof. According to Theorem 5.2.5 $\breve{\beta} \xrightarrow{P} \beta$. So it suffices to show that $\tilde{\beta} - \breve{\beta} \xrightarrow{P} 0$. We use an induction argument. For $i = 1$, GLS and EGLS estimation are equivalent because they both coincide with OLS estimation. For general $i(= 2, \dots, r)$,

$$\begin{aligned}\tilde{\beta}_i - \breve{\beta}_i &\overset{1}{=} G_i X_i'(\breve{\zeta}_i - \tilde{\zeta}_i) \overset{2}{=} G_i X_i' \left(Y_{(i-1)}(\alpha_i - \tilde{\alpha}_{0i}) + \tilde{\mu}_{(i-1)}\tilde{\alpha}_{0i} - \breve{\mu}_{(i-1)}\alpha_i\right) \\ &\overset{3}{=} G_i X_i' Y_{(i-1)}(\alpha_i - \tilde{\alpha}_{0i}) + \tilde{\beta}_{(i-1)}(\tilde{\alpha}_{0i} - \alpha_i) + (\tilde{\beta}_{(i-1)} - \breve{\beta}_{(i-1)})\alpha_i.\end{aligned}$$

The first equality follows from the definitions of the (E)GLS estimators (4.3.16) and (5.2.12). The second equation follows by definition from $\breve{\zeta}_i = \breve{\varepsilon}_{(i-1)}\alpha_i$ and $\breve{\varepsilon}_{(i-1)} = Y_{(i-1)} - \breve{\mu}_{(i-1)}$ (and similarly for $\tilde{\zeta}_i$). Rewriting gives the third equation.

Note that $G_i X_i' Y_{(i-1)}$ can be considered as an OLS estimator for $\beta_{(i-1)}$ based on the first N_i observations. Since $N_i \geq N_r \to \infty$, a similar proof as for Theorem 5.2.3 gives $G_i X_i' Y_{(i-1)} \xrightarrow{P} \beta_{(i-1)}$. All three terms converge in probability to zero because of the condition $\tilde{\alpha}_{0i} \xrightarrow{P} \alpha_i$ and the induction assumption $\tilde{\beta}_{(i-1)} - \breve{\beta}_{(i-1)} \xrightarrow{P} 0$. □

To prove consistency of EGLS estimators for Σ we need the following Lemma.

Lemma 5.2.7. *If* $\tilde{\alpha}_{0i} \xrightarrow{P} \alpha_i$ *for* $i = 1, \ldots, r$, *then* $\tilde{\Gamma}_{ii} \xrightarrow{P} \Gamma_{ii}$.

Proof. We have

$$\begin{aligned}\frac{1}{N_i}|U_i\varepsilon_{(i-1)}(\tilde{\alpha}_{0i} - \alpha_i)|^2 &\leq \frac{1}{N_i}|U_i\varepsilon_{(i-1)}(\tilde{\alpha}_{0i} - \alpha_i)|^2 + \frac{1}{N_i}|H_i\varepsilon_{(i-1)}(\tilde{\alpha}_{0i} - \alpha_i)|^2 \\ &= \frac{1}{N_i}(\tilde{\alpha}_{0i} - \alpha_i)'\varepsilon'_{(i-1)}\varepsilon_{(i-1)}(\tilde{\alpha}_{0i} - \alpha_i) \xrightarrow{P} 0.\end{aligned}$$

The equality follows from $I_{N_i} = U_i + H_i$, the convergence in probability from (5.2.4) and the condition $\tilde{\alpha}_{0i} \xrightarrow{P} \alpha_i$.

We have

$$\begin{aligned}\tilde{\eta}_i &\overset{1}{=} U_i(Y_i - \varepsilon_{(i-1)}\tilde{\alpha}_{0i}) \overset{2}{=} U_i(Y_i - \mu_i - \varepsilon_{(i-1)}\alpha_i + \varepsilon_{(i-1)}(\alpha_i - \tilde{\alpha}_{0i})) \\ &\overset{3}{=} U_i(\eta_i + \varepsilon_{(i-1)}(\alpha_i - \tilde{\alpha}_{0i})).\end{aligned}$$

An argumentation as in the proof of Theorem 4.3.2 leads to the first equality. Since $\mu_i \in L_i$, we have $U_i\mu_i = 0$ and the second equality holds. The third equation follows from (4.3.12).

As in the proof of Lemma 5.2.1 we have $|H_i\eta_i| = O_P(1)$. Combining this with the two previous results gives

$$|\eta_i - \tilde{\eta}_i| = |\eta_i - U_i\eta_i - U_i\varepsilon_{(i-1)}(\alpha_i - \tilde{\alpha}_{0i})| \leq |H_i\eta_i| + |U_i\varepsilon_{(i-1)}(\alpha_i - \tilde{\alpha}_{0i})| = o_P(N_r^{\frac{1}{2}}).$$

According to (5.2.8) $|\eta_i| = O_P(N_r^{\frac{1}{2}})$ so that $|\tilde{\eta}_i| \leq |\tilde{\eta}_i - \eta_i| + |\eta_i| = O_P(N_r^{\frac{1}{2}})$ as well. This gives

$$|\eta_i - \tilde{\eta}_i|(|\eta_i| + |\tilde{\eta}_i|) = o_P(N_r).$$

Since $|N_i\tilde{\Gamma}_{ii} - \eta_i\eta_i| = |\tilde{\eta}_i'\tilde{\eta}_i - \eta_i\eta_i| \leq |\eta_i - \tilde{\eta}_i|(|\eta_i| + |\tilde{\eta}_i|)$ and $\frac{1}{N_i}\eta_i'\eta_i \xrightarrow{P} \Gamma_{ii}$ according to (5.2.5), this proves the lemma. □

Similar to OLS estimation, we define a class of EGLS estimators for Σ in which each estimator has entries

$$\tilde{S}_{ig}^{(h)} = \frac{\tilde{\varepsilon}_i^{(h)'}\tilde{\varepsilon}_g^{(h)}}{N_h}, \quad g = 1, \ldots, i, \quad i = 1, \ldots, r, \quad \text{with } h \in \{i, \ldots, r\}.$$

Theorem 5.2.8. *If* $\tilde{\alpha}_{0i} \xrightarrow{P} \alpha_i$ *for* $i = 1, \ldots, r$, *then* $\tilde{S}_{ig}^{(h)} \xrightarrow{P} \sigma_{ig}$.

Proof. We prove the theorem by induction. For $i = 1$, EGLS and OLS estimation coincide and $\tilde{S}_{11}^{(h)} \xrightarrow{P} \sigma_{11}$ according to Theorem 5.2.4. For general i $(= 2, \ldots, r)$, the induction assumption implies that $\frac{1}{N_i}\tilde{\varepsilon}'_{(i-1)}\tilde{\varepsilon}_{(i-1)} \xrightarrow{P} \Sigma_{(i-1)(i-1)}$. In combination with the condition $\tilde{\alpha}_{0i} \xrightarrow{P} \alpha_i$, this leads to the following convergence in probability

$$\frac{1}{N_i}\tilde{\zeta}_i'\tilde{\zeta}_i = \frac{1}{N_i}\tilde{\alpha}'_{0i}\tilde{\varepsilon}'_{(i-1)}\tilde{\varepsilon}_{(i-1)}\tilde{\alpha}_{0i} \xrightarrow{P} \alpha_i\Sigma_{(i-1)(i-1)}\alpha_i = \Delta_{ii}, \tag{5.2.13}$$

where the first equality follows from $\tilde{\zeta}_i = \tilde{\varepsilon}_{(i-1)}\tilde{\alpha}_{0i}$ (similar to (4.3.12)), and the last equality follows from (4.3.9).

According to Lemma 5.2.7 we have the following convergence in probability

$$\frac{1}{N_i}\tilde{\eta}_i'\tilde{\eta}_i = \frac{1}{N_i}(\tilde{\varepsilon}_i'\tilde{\varepsilon}_i - \tilde{\zeta}_i'\tilde{\zeta}_i) \xrightarrow{P} \Gamma_{ii} = \Sigma_{ii} - \Delta_{ii}, \tag{5.2.14}$$

where the first equality follows from $\tilde{\varepsilon}_i = \tilde{\zeta}_i + \tilde{\eta}_i$ and the orthogonality of $\tilde{\zeta}_i$ and $\tilde{\eta}_i$, and the last equality by definition from (4.3.9). Combining (5.2.13) and (5.2.14) gives

$$\frac{1}{N_i}\tilde{\varepsilon}_i'\tilde{\varepsilon}_i \xrightarrow{P} \Sigma_{ii}. \tag{5.2.15}$$

For every matrix Σ relation (4.3.15) holds, so also for $\tilde{S}_0$:

$$\tilde{\varepsilon}_i = E_i + H_i\tilde{\zeta}_i. \tag{5.2.16}$$

Since E_i and $H_i\tilde{\zeta}_i$ are orthogonal we have $\tilde{\varepsilon}_i'\tilde{\varepsilon}_i = E_i'E_i + (H_i\tilde{\zeta}_i)'(H_i\tilde{\zeta}_i)$. Substituting this in (5.2.15) gives $\frac{1}{N_i}(E_i'E_i + (H_i\tilde{\zeta}_i)'(H_i\tilde{\zeta}_i)) \xrightarrow{P} \Sigma_{ii}$. According to Theorem 5.2.4 $\frac{1}{N_i}E_i'E_i \xrightarrow{P} \Sigma_{ii}$, so $|H_i\tilde{\zeta}_i| = o_P(N_r^{\frac{1}{2}})$. We have

$$\begin{aligned}
|\tilde{\varepsilon}_i^{(h)'}\tilde{\varepsilon}_g^{(h)} - \varepsilon_i^{(h)'}\varepsilon_g^{(h)}| &\overset{1}{\leq} |E_i^{(h)'}E_g^{(h)} - \varepsilon_i^{(h)'}\varepsilon_g^{(h)}| + |E_i^{(h)}||(H_g\tilde{\zeta}_g)^{(h)}| \\
&\quad + |(H_i\tilde{\zeta}_i)^{(h)}||E_g^{(h)}| + |(H_i\tilde{\zeta}_i)^{(h)}||(H_g\tilde{\zeta}_g)^{(h)}| \\
&\overset{2}{\leq} |E_i^{(h)}||E_g^{(h)} - \varepsilon_g^{(h)}| + |E_i^{(h)} - \varepsilon_i^{(h)}||\varepsilon_g^{(h)}| \\
&\quad + |E_i^{(h)}||H_g\tilde{\zeta}_g| + |H_i\tilde{\zeta}_i||E_g^{(h)}| + |H_i\tilde{\zeta}_i||H_g\tilde{\zeta}_g| \\
&\overset{3}{=} o_P(N_r),
\end{aligned}$$

where we used (5.2.16) and (5.2.9) to obtain the first and second inequality, respectively. The third relation follows from $|H_i\tilde{\zeta}_i| = o_P(N_r^{\frac{1}{2}})$, (5.2.10) and (5.2.11). Since $\frac{1}{N_h}\varepsilon_i^{(h)'}\varepsilon_g^{(h)} \xrightarrow{P} \sigma_{ig}$ according to (5.2.4), this completes the proof. □

Similar to $\widehat{S}$ in (4.3.21), we construct $\tilde{S}$ as

$$\begin{cases} \tilde{S}_{ii} = \tilde{\varepsilon}_i'\tilde{\varepsilon}_i/r_i \\ \tilde{S}_{ig} = \tilde{\varepsilon}_i'\tilde{\varepsilon}_{(i-1)g}/r_i \quad \text{for } g = 1, ..., i-1, \end{cases}$$

and based on (4.3.8) we construct

$$\tilde{\alpha}_i = \tilde{S}_{(i-1)(i-1)}^{-1}\tilde{S}_{(i-1)i}. \tag{5.2.17}$$

Similar to the MLE $\widehat{\Sigma}$ in (4.3.26) we construct $\tilde{\Sigma}$ as

$$\begin{cases} \tilde{\Sigma}_{11} = \tilde{\Gamma}_{11} \text{ and for } i = 2, ..., r\text{:} \\ \tilde{\Sigma}_{(i-1)i} = \tilde{\Sigma}_{(i-1)(i-1)}\tilde{\alpha}_i,\ \tilde{\Delta}_{ii} = \tilde{\alpha}_i'\tilde{\Sigma}_{(i-1)(i-1)}\tilde{\alpha}_i,\ \tilde{\Sigma}_{ii} = \tilde{\Gamma}_{ii} + \tilde{\Delta}_{ii}. \end{cases} \tag{5.2.18}$$

Theorem 5.2.9. *If $\tilde{\alpha}_{0i} \xrightarrow{P} \alpha_i$ for $i = 1, \ldots, r$, then $\tilde{S} \xrightarrow{P} \Sigma$ and $\tilde{\Sigma} \xrightarrow{P} \Sigma$.*

Proof. Since $\tilde{S}_{ig} = \frac{N_i}{r_i}\tilde{S}_{ig}^{(i)}$ and $\frac{N_i}{r_i} \to 1$, the consistency of $\tilde{S}$ follows directly from Theorem 5.2.8.

The α_i in (4.3.8) are continuous function of Σ. Since the $\tilde{\alpha}_i$ in (5.2.17) are the same continuous functions of consistent $\tilde{S}$, the $\tilde{\alpha}_i$ are consistent as well. In combination with Lemma 5.2.7 this proves the consistency of $\tilde{\Sigma}$ in (5.2.18) since $\tilde{\Sigma}$ is the same continuous function of $\tilde{\Gamma}_{ii}$ and $\tilde{\alpha}_i$ as Σ is of Γ_{ii} and α_i. □

According to Theorems 5.2.6 and 5.2.9, a consistent starting estimator $\tilde{S}_0$ (and consequently consistent $\tilde{\alpha}_{0i}$) results in consistent EGLS estimators. In practice it is common to perform OLS estimation and then to take the resulting OLS estimator S as $\tilde{S}_0$. Since S is consistent according to Corollary 5.2.4, this results in consistent EGLS estimators.

In iterative EGLS, the EGLS estimation procedure is repeated several times and the estimate for Σ of an iteration is taken as the starting estimate in the next iteration. For our model, it is clear that such an iterative procedure would result in consistent estimators in each step, if the initial estimator S_0 is consistent.

5.2.5 ML

In Section 4.3.5, the MLE's were derived in case all errors are normally distributed: from general theory it is known that these MLE's are consistent under certain regularity conditions. Here we will prove consistency if the normality assumption is dropped.

In Section 4.3.5 we have seen that ML estimation coincides with a specific type of EGLS estimation. The estimators $\widehat{\alpha}_i$ and $\widehat{\beta}$ in (4.3.20) were derived by simultaneously minimizing the GLS criterium w.r.t. α_i and β. However, $\widehat{\beta}$ (and consequently $\widehat{S}$ in (4.3.21) and $\widehat{\Sigma}$ in (4.3.26)) can also be derived by means of EGLS estimation with $\widehat{\alpha}_i$ as starting value $\tilde{\alpha}_{0i}$. For this, a closed form expression for $\widehat{\alpha}_i$ is required, which we derive by means of partial regression. In the first step we regress Y_i and $\widehat{\varepsilon}_{(i-1)}$ onto X_i. In the second step we regress the residuals of Y_i onto the residuals of $\widehat{\varepsilon}_{(i-1)}$. This results in

$$\widehat{\alpha}_i = (\widehat{\varepsilon}'_{(i-1)} U_i \widehat{\varepsilon}_{(i-1)})^{-1} \widehat{\varepsilon}'_{(i-1)} U_i Y_i = (\varepsilon'_{(i-1)} U_i \varepsilon_{(i-1)})^{-1} \varepsilon'_{(i-1)} U_i \varepsilon_i. \quad (5.2.19)$$

Theorem 5.2.10. $\widehat{\alpha}_i \xrightarrow{P} \alpha_i$, $\widehat{\beta} \xrightarrow{P} \beta$, $\widehat{\Gamma}_{ii} \xrightarrow{P} \Gamma_{ii}$, $\widehat{S} \xrightarrow{P} \Sigma$ *and* $\widehat{\Sigma} \xrightarrow{P} \Sigma$.

Proof. We denote $\varepsilon = [\varepsilon_{(i-1)}\ \varepsilon_i]$. Since $\varepsilon' H_i \varepsilon \geq 0$ and

$$tr\left(E\{\varepsilon' H_i \varepsilon\}\right) = tr\left(H_i E\{\varepsilon\varepsilon'\}\right) = tr(H_i) \sum_{g=1}^{i} \sigma_{gg} \leq k \sum_{g=1}^{i} \sigma_{gg},$$

we see that $\frac{1}{N_i}(\varepsilon'\varepsilon - \varepsilon' U_i \varepsilon) = \frac{1}{N_i}\varepsilon' H_i \varepsilon \xrightarrow{P} 0$. In combination with (5.2.4), this gives

$$\begin{aligned}\widehat{\alpha}_i - (\varepsilon'_{(i-1)}\varepsilon_{(i-1)})^{-1}\varepsilon'_{(i-1)}\varepsilon_i = &-(\frac{1}{N_i}\varepsilon'_{(i-1)}U_i\varepsilon_{(i-1)})^{-1}\frac{1}{N_i}\varepsilon'_{(i-1)}H_i\varepsilon_i \\ &+\left((\frac{1}{N_i}\varepsilon'_{(i-1)}U_i\varepsilon_{(i-1)})^{-1} - (\frac{1}{N_i}\varepsilon'_{(i-1)}\varepsilon_{(i-1)})^{-1}\right)\frac{1}{N_i}\varepsilon'_{(i-1)}\varepsilon_i \xrightarrow{P} 0.\end{aligned}$$

From (5.2.4) we also get

$$(\frac{1}{N_i}\varepsilon'_{(i-1)}\varepsilon_{(i-1)})^{-1}\frac{1}{N_i}\varepsilon'_{(i-1)}\varepsilon_i \xrightarrow{P} \Sigma^{-1}_{(i-1)(i-1)}\Sigma_{(i-1)i} = \alpha_i.$$

Together this implies

$$\widehat{\alpha}_i - \alpha_i = \left(\widehat{\alpha}_i - (\varepsilon'_{(i-1)}\varepsilon_{(i-1)})^{-1}\varepsilon'_{(i-1)}\varepsilon_i\right) + \left((\varepsilon'_{(i-1)}\varepsilon_{(i-1)})^{-1}\varepsilon'_{(i-1)}\varepsilon_i - \alpha_i\right) \xrightarrow{P} 0.$$

Since ML estimation is a specific kind of EGLS estimation, all convergence properties of Section 5.2.4 still hold. Accordingly, the MLE's (for β, Γ_{ii} and Σ) are consistent if the $\widehat{\alpha}_i (= \tilde{\alpha}_{0i})$ are consistent. □

5.3 Iterative EGLS

5.3.1 Introduction

In this section we look in more detail at the iterative EGLS procedure and the properties of the estimators in each iteration. We consider the specific EGLS procedure where in each iteration the estimators for β and Σ are the conditional MLE's under the normality assumption in the following sense. Each iteration consists of two steps: first the ML estimate for β is determined given a previously determined estimate for Σ, secondly the ML estimate for Σ is determined given the previous estimate for β.

There are different ways to determine these conditional estimators. Srivastava (1985) used matrix differentiation to derive the first order conditions for multivariate regression with a *general* missing data pattern. These first order conditions can also be used for a monotone missing data pattern. However, they consist of non-linear matrix equations which have to be solved numerically. For the numerical example of Chapter 4 (which has a small number of observations), this caused problems for the iterative algorithms which we used.

In order to construct the EGLS algorithm in an alternative way, we first discuss ML estimation of Σ with known regression coefficients in Section 5.3.2. This technique is used in the iterative EGLS procedure which is presented in Section 5.3.3.

5.3.2 ML estimation of Σ with known β

We assume the model of Section 4.3.5 but with known regression coefficients β. Similar to Chapter 4, the MLE's are derived by means of orthogonal projections. We introduce the following additional notation

$$\begin{array}{lcll} L_{\varepsilon(i)} & = & \mathcal{R}(\varepsilon_{(i-1)}), & \\ H_{\varepsilon(i)} & \in & \mathbb{R}^{N_i \times N_i} & \text{: orthogonal projection matrix of } L_{\varepsilon(i)}, \\ U_{\varepsilon(i)} & = & I_{N_i} - H_{\varepsilon(i)} & \text{: orthogonal projection matrix of } L_{\varepsilon(i)}^{\perp}. \end{array}$$

Theorem 5.3.1. *The MLE for α_i is*

$$\breve{\alpha}_i = (\varepsilon'_{(i-1)}\varepsilon_{(i-1)})^{-1}\varepsilon_{(i-1)}\varepsilon_i,$$

and the MLE for Γ_{ii} is

$$\breve{\Gamma}_{ii} = \frac{\varepsilon_i' U_{\varepsilon(i)}\varepsilon_i}{N_i}. \tag{5.3.1}$$

Proof. The likelihood reads

$$\begin{aligned} L(\Sigma;\beta,Y) &\stackrel{1}{=} \prod_{i=1}^{r}\Big[\{(2\pi)^{m_i}|\Gamma_{ii}|\}^{-\frac{N_i}{2}}\exp\{-\tfrac{1}{2}tr(\Gamma_{ii}^{-1}\eta_i'\eta_i\}\Big] \\ &\stackrel{2}{=} \prod_{i=1}^{r}\Big[\{(2\pi)^{m_i}|\Gamma_{ii}|\}^{-\frac{N_i}{2}}\exp\{-\tfrac{1}{2}tr(\Gamma_{ii}^{-1}(\eta_i'U_{\varepsilon(i)}\eta_i + \eta_i'H_{\varepsilon(i)}\eta_i)\}\Big] \\ &\stackrel{3}{=} \prod_{i=1}^{r}\Big[\{(2\pi)^{m_i}|\Gamma_{ii}|\}^{-\frac{N_i}{2}}\exp\{-\tfrac{1}{2}tr(\Gamma_{ii}^{-1}\varepsilon_i'U_{\varepsilon(i)}\varepsilon_i + \\ &\qquad \Gamma_{ii}^{-1}(Y_i-\mu_i-\varepsilon_{(i-1)}\alpha_i)'H_{\varepsilon(i)}(Y_i-\mu_i-\varepsilon_{(i-1)}\alpha_i))\}\Big]. \end{aligned} \tag{5.3.2}$$

See (4.3.22) for the first equality. The second equality holds because the projection matrices $H_{\varepsilon(i)}$ and $U_{\varepsilon(i)}$ are orthogonal and $H_{\varepsilon(i)} + U_{\varepsilon(i)} = I_{N_i}$. The third equality follows from $\eta_i = Y_i - \mu_i - \varepsilon_{(i-1)}\alpha_i$, (4.3.12) and $U_{\varepsilon(i)}\eta_i = U_{\varepsilon(i)}\varepsilon_i$ (because $\varepsilon_{(i-1)}\alpha_i \in L_{\varepsilon(i)}$ and thus $U_{\varepsilon(i)}\varepsilon_{(i-1)}\alpha_i = 0$).

The MLE's are obtained by maximization of (5.3.2) w.r.t. all α_i and Γ_{ii}, respectively. Regardless of the value of Γ_{ii}, the term $H_{\varepsilon(i)}(Y_i-\mu_i-\varepsilon_{(i-1)}\alpha_i)$ is zero for $\breve{\alpha}_i = (\varepsilon'_{(i-1)}\varepsilon_{(i-1)})^{-1}\varepsilon_{(i-1)}\varepsilon_i$. Substitution of $\breve{\alpha}_i$ in (5.3.2) gives

$$\sup_{\alpha_i} L(\Sigma;\beta,Y) = \prod_{i=1}^{r}[\{(2\pi)^{m_i}|\Gamma_{ii}|\}^{-\frac{N_i}{2}}\exp\{-\tfrac{1}{2}tr(\Gamma_{ii}^{-1}\varepsilon_i'U_{\varepsilon(i)}\varepsilon_i)\}].$$

A similar reasoning as in the proof of Theorem 4.3.5 leads to the MLE for Γ_{ii}. □

The MLE $\breve{\Sigma}$ for the covariance matrix follows sequentially from the relations:

$$\begin{cases} \breve{\Sigma}_{11} = \breve{\Gamma}_{11} \text{ and for } i = 2, ..., r: \\ \breve{\Sigma}_{(i-1)i} = \breve{\Sigma}_{(i-1)(i-1)}\breve{\alpha}_i, \ \breve{\Delta}_{ii} = \breve{\alpha}_i'\breve{\Sigma}_{(i-1)(i-1)}\breve{\alpha}_i, \ \breve{\Sigma}_{ii} = \breve{\Gamma}_{ii} + \breve{\Delta}_{ii}. \end{cases} \tag{5.3.3}$$

Theorem 5.3.2. $\breve{\Sigma} \xrightarrow{P} \Sigma$.

Proof. We have

$$\eta_i'\eta_i - N_i\breve{\Gamma}_{ii} \overset{1}{=} \eta_i'\eta_i - \varepsilon_i'U_{\varepsilon(i)}\varepsilon_i \overset{2}{=} \eta_i'\eta_i - \eta_i'U_{\varepsilon(i)}\eta_i \overset{3}{=} \eta_i'H_{\varepsilon(i)}\eta_i \overset{4}{=} o_P(N_r).$$

The first equality follows by definition from (5.3.1), the second from $\varepsilon_i = \varepsilon_{(i-1)}\alpha_i + \eta_i$ and $U_{\varepsilon(i)}\varepsilon_{(i-1)} = 0$, the third from $I_{N_i} = H_{\varepsilon(i)} + U_{\varepsilon(i)}$. The fourth relation follows from $\eta_i'H_{\varepsilon(i)}\eta_i \geq 0$ and $tr(E\{\frac{1}{N_i}\eta_i'H_{\varepsilon(i)}\eta_i\}) \leq \frac{M_{i-1}\Gamma_{ii}}{N_i} \to 0$.

Since $\frac{1}{N_i}\eta_i'\eta_i \xrightarrow{P} \Gamma_{ii}$ according to (5.2.5), this proves $\breve{\Gamma}_{ii} \xrightarrow{P} \Gamma_{ii}$. From (5.2.4) it follows that $\breve{\alpha}_i \xrightarrow{P} \alpha_i$. Since $\breve{\Sigma}$ in (5.3.3) is the same continuous function of $\breve{\alpha}_i$ and $\breve{\Gamma}_{ii}$ as Σ is of α_i and Γ_{ii}, this completes the proof. □

5.3.3 The iterative EGLS procedure

In each iteration estimates for β and Σ have to be determined, or equivalently, the estimates for β, α_i and Γ_{ii} have to be determined. In the procedure we discuss here, the estimates in iteration q, $\breve{\beta}_q$, $\breve{\alpha}_{qi}$ and $\breve{\Gamma}_{qii}$, are the conditional ML estimates under the normality assumption. So $\breve{\beta}_{qi}$ is the EGLS estimator for β with starting value $\breve{\alpha}_{q-1,i}$ (see (5.2.12)). Similarly, $\breve{\alpha}_{qi}$ and $\breve{\Gamma}_{qii}$ are the MLE's for α_i and Γ_{ii} given $\beta = \breve{\beta}_q$ (see Theorem 5.3.1). Summarized, iteration q of the iterative EGLS procedure consists of the follow three steps:

(i) $\breve{\beta}_{qi} = G_iX_i'(Y_i - \breve{\varepsilon}_{q-1,(i-1)}\breve{\alpha}_{q-1,i})$

(ii) $\breve{\alpha}_{qi} = (\breve{\varepsilon}'_{q(i-1)}\breve{\varepsilon}_{q(i-1)})^{-1}\breve{\varepsilon}'_{q(i-1)}\breve{\varepsilon}_{qi}$

(iii) $\breve{\Gamma}_{qii} = \frac{1}{N_i}\breve{\varepsilon}'_{qi}U_{\breve{\varepsilon}_q(i)}\breve{\varepsilon}_{qi}$

where

$$\begin{array}{lll} \breve{\varepsilon}_{qi} & = Y_i - X_i\breve{\beta}_{qi}, & \\ H_{\breve{\varepsilon}_q(i)} & \in \mathbb{R}^{N_i \times N_i}, & \text{orthogonal projection matrix of } \mathcal{R}(\breve{\varepsilon}_{q(i-1)}), \\ U_{\breve{\varepsilon}_q(i)} & = I_{N_i} - H_{\breve{\varepsilon}_q(i)}, & \text{orthogonal projection matrix of } \mathcal{R}(\breve{\varepsilon}_{q(i-1)})^{\perp}. \end{array}$$

Step (iii) could be omitted from the iterative procedure, because only $\breve{\alpha}_{qi}$ is used in the next iteration and not $\breve{\Gamma}_{qii}$. Only in the last iteration, step (iii) needs to be executed to determine the final estimate for Σ.

Similar to the MLE $\widehat{\Sigma}$ in (4.3.26) we construct the EGLS estimate $\tilde{\Sigma}_q$ in iteration q as

$$\begin{cases} \tilde{\Sigma}_{q11} = \tilde{\Gamma}_{q11} \text{ and for } i = 2, ..., r: \quad \tilde{\Sigma}_{q(i-1)i} = \tilde{\Sigma}_{q(i-1)(i-1)}\tilde{\alpha}_{qi}, \\ \tilde{\Delta}_{qii} = \tilde{\alpha}'_{qi}\tilde{\Sigma}_{q(i-1)(i-1)}\tilde{\alpha}_{qi}, \ \tilde{\Sigma}_{qii} = \tilde{\Gamma}_{qii} + \tilde{\Delta}_{qii}. \end{cases} \tag{5.3.4}$$

Theorem 5.3.3. *If $\breve{\alpha}_{0i} \xrightarrow{P} \alpha_i$ and $\|X_i'X_i\| = O(N_r)$, then the estimators are consistent in each iteration, in particular:*

$$\begin{aligned} \breve{\beta}_q &\xrightarrow{P} \beta, \\ \breve{\Sigma}_q &\xrightarrow{P} \Sigma. \end{aligned}$$

Proof. Without loss of generality we take $m_i = 1$. The consistency of $\breve{\beta}_q$ follows directly from Theorem 5.2.6. As a consequence of this consistency and the condition $\|X_i'X_i\| = O(N_r)$, we have $|\breve{\mu}_{qi} - \mu_i|^2 = (\breve{\beta}_{qi} - \beta_i)'X_i'X_i(\breve{\beta}_{qi} - \beta_i) = o_p(N_r)$. Hence, $|\breve{\varepsilon}_{qi} - \varepsilon_i| = |\breve{\mu}_{qi} - \mu_i| = o_p(N_r^{\frac{1}{2}})$. In combination with (5.2.4), this proves the consistency of $\breve{\alpha}_{qi}$.

From the consistency of $\breve{\alpha}_{qi}$ and $|\breve{\varepsilon}_{qi} - \varepsilon_i| = o_P(N_r^{\frac{1}{2}})$, it follows that $|\breve{\varepsilon}_{q(i-1)}(\breve{\alpha}_{qi} - \alpha_i)| = o_P(N_r^{\frac{1}{2}})$ and $|(\breve{\varepsilon}_{q(i-1)} - \varepsilon_{(i-1)})\alpha_i| = o_P(N_r^{\frac{1}{2}})$. Hence

$$|\breve{\varepsilon}_{q(i-1)}\breve{\alpha}_{qi} - \varepsilon_{(i-1)}\alpha_i| \leq |\breve{\varepsilon}_{q(i-1)}(\breve{\alpha}_{qi} - \alpha_i)| + |(\breve{\varepsilon}_{q(i-1)} - \varepsilon_{(i-1)})\alpha_i| = o_P(N_r^{\frac{1}{2}}).$$

Similarly we have

$$|U_{\breve{\varepsilon}_q(i)}(\breve{\eta}_{qi} - \eta_i)| \leq |\breve{\eta}_{qi} - \eta_i| \leq |\breve{\mu}_{qi} - \mu_i| + |\breve{\varepsilon}_{q(i-1)}\breve{\alpha}_{qi} - \varepsilon_{(i-1)}\alpha_i| = o_P(N_r^{\frac{1}{2}}),$$

where the first inequality follows from $U_{\breve{\varepsilon}_{q(i)}} \leq I_{N_i}$. The second inequality follows from $\breve{\eta}_{qi} = Y_i - \breve{\mu}_{qi} - \breve{\varepsilon}_{q(i-1)}\breve{\alpha}_{qi}$ and (4.3.12).

Since $tr(E\{\eta'_{qi}H_{\breve{\varepsilon}_{q(i)}}\eta_{qi}|Y_{(i-1)}\}) \leq M_{i-1}\Gamma_{ii}$ is finite, we have $|H_{\breve{\varepsilon}_{q(i)}}\eta_{qi}| = o_P(N_r^{\frac{1}{2}})$. So

$$|\eta_i - U_{\breve{\varepsilon}_{q(i)}}\breve{\eta}_{qi}| \leq |U_{\breve{\varepsilon}_{q(i)}}(\eta_i - \breve{\eta}_{qi})| + |H_{\breve{\varepsilon}_{q(i)}}\eta_{qi}| = o_P(N_r^{\frac{1}{2}}). \tag{5.3.5}$$

According to (5.2.8) $|\eta_i| = O_P(N_r^{\frac{1}{2}})$, so

$$|U_{\breve{\varepsilon}_{q(i)}}\breve{\eta}_{qi}| \leq |\eta_i - U_{\breve{\varepsilon}_{q(i)}}\breve{\eta}_{qi}| + |\eta_i| = O_P(N_r^{\frac{1}{2}}). \tag{5.3.6}$$

We have

$$\begin{aligned} |\eta'_i\eta_i - N_i\breve{\Gamma}_{qii}| &= |\eta'_i\eta_i - \breve{\eta}'_{qi}U_{\breve{\varepsilon}_{q(i)}}\breve{\eta}_{qi}| \\ &\leq |\eta_i - U_{\breve{\varepsilon}_{q(i)}}\breve{\eta}_{qi}|(|\eta_i| + |U_{\breve{\varepsilon}_{q(i)}}\breve{\eta}_{qi}|) = o_P(N_r), \end{aligned}$$

where the first equality follows from (5.3.1) and $U_{\breve{\varepsilon}_{q(i)}}\breve{\varepsilon}_i = U_{\breve{\varepsilon}_{q(i)}}\breve{\eta}_i$. The second equality follows from $|\eta_i| = O_P(N_r^{\frac{1}{2}})$, (5.3.5) and (5.3.6).

This proves that Γ_{ii} is consistent because $\frac{1}{N_i}\eta'_i\eta_i \xrightarrow{P} \Gamma_{ii}$ according to (5.2.5). Since $\breve{\Sigma}_q$ in (5.3.4) is the same continuous function of $\breve{\Gamma}_{qii}$ and $\breve{\alpha}_{qi}$ as Σ is of α_i and Γ_{ii}, this completes the proof. □

Though the estimators in the iterative EGLS procedure are consistent in each iteration, this does not necessarily mean that they share the same asymptotic properties of the MLE's, such as asymptotic efficiency. See for precise conditions Magnus (1978), Theorem 4. We leave the verification of these conditions for further research.

Numerical illustration

We applied the described iterative EGLS algorithm to the numerical example of Section 4.2. As starting value $\breve{\Sigma}_0$ we took the OLS estimate S. The algorithm only needed 3 (5) iterations to produce the maximum likelihood estimates, accurate up to two (four) decimals. However, the numerical example concerns only

a small number of observations with a relatively low fraction of missing observations. In practical problems the iterative EGLS algorithm will obviously need more iterations to converge.

5.4 EM-algorithm

5.4.1 Introduction

The EM-algorithm (and generalizations of it, such as the ECM-algorithm, see McLachan and Krishnan (1997) for an overview) is a widely used technique in missing data problems to determine ML estimates. The EM-algorithm is an iterative procedure which has been proven to converge numerically to the ML estimates under certain conditions (see Dempster *et al*. (1977) and Wu (1983) *e.g.*). In this section we look in more detail at the EM-algorithm for the model of Section 4.3.5, *i.e.* for the multivariate regression model with monotone missing data of the dependent variables and normally distributed errors. We will also give the EM-algorithm for a general missing data pattern.

The underlying idea of the EM-algorithm is that it might be difficult to determine the MLE's from the observed (incomplete) data, but it would be simple in case of complete data. Therefore the missing observations are substituted by their expected values and subsequently the ML estimates are determined from the completed data. Based on these new estimates, the expected values of the missing observations are again determined, *et cetera*. Accordingly, each iteration of the EM-algorithm consists of an E(xpectation) and a M(aximization) step.

In Meng and Rubin (1993) the ECM-algorithm was presented for a multivariate regression model which is similar to our model but differs in two aspects:

1. the explanatory variables do not necessarily have identical values for all dependent variables,

2. the regression coefficients are identical for all dependent variables ($\beta_1 = \ldots = \beta_r$).

Our model considers the special case of identical explanatory variables for all dependent variables. As a consequence of the identical explanatory variables the M-step can be simplified considerably.

5.4.2 Additional notation

To describe the EM-algorithm for a general missing data pattern, we need the following additional notation:

obs_t : set of indices of the groups of dependent variables for which the observations are present for case t,

mis_t : set of indices of the groups of dependent variables for which the observations are missing for case t,

obs : set of indices of the observed values of the groups of dependent variables,

mis : set of indices of the missing values of the groups of dependent variables,

Y = (Y_{obs}, Y_{mis})
: matrix of all (observed and unobserved) values of the dependent variables,

X_t $\in \mathbb{R}^{k\times 1}$
: values of the explanatory variables for observation t,

$\Sigma_{mis_t mis_t \cdot obs_t}$ = $\Sigma_{mis_t mis_t} - \Sigma_{mis_t obs_t} \left(\Sigma_{obs_t obs_t}\right)^{-1} \Sigma_{obs_t mis_t}$
: conditional variance of the missing variables given the observed variables for case t.

We will denote the estimators for the parameters in iteration q of the EM-algorithm by the corresponding symbols plus a superscript $\widehat{\ }$(similar to the MLE's) and an additional subindex q.

5.4.3 E-step

In the Expectation step, the expectation of the sufficient statistics is calculated, given the (estimated) values of the parameters characterizing the complete data likelihood. For our model, this comes down to determine the expectations of the missing values themselves and their cross-products.

E-step: general missing data pattern
Expectations missing values:

$$E\{Y_{ti}|Y_{obs}, \widehat{\beta}_q, \widehat{\Sigma}_q\} = Y_{qti},$$

where

$$\begin{aligned} Y_{qti} &= \begin{cases} Y_{ti}, & i \in obs_t \\ \widehat{\nu}_{qti}, & i \in mis_t \end{cases} \\ &= \begin{cases} Y_{ti}, & i \in obs_t \\ X_t'\widehat{\beta}_{qi} + \widehat{\Sigma}_{qi,obs_t}\widehat{\Sigma}^{-1}_{q,obs_t obs_t}\left(Y_{qt,obs_t} - \widehat{\beta}'_{q,obs_t}X_t\right), & i \in mis_t. \end{cases} \end{aligned}$$

Expectations inner products missing values:

$$E\{Y_{ti}Y'_{tj}|Y_{obs}, \widehat{\beta}_q, \widehat{\Sigma}_q\} = Y_{qti}\,(Y_{qtj})' + c_{qtij}, \text{ where}$$

$$c_{qtij} = \begin{cases} 0, & i \in obs_t \text{ and/or } j \in obs_t \\ \widehat{\Sigma}_{qij\cdot obs_t}, & i \in mis_t \text{ and } j \in mis_t, \end{cases}$$

with $\widehat{\Sigma}_{qij\cdot obs_t}$, the appropriate elements of $\widehat{\Sigma}_{q,mis_t mis_t\cdot obs_t}$.

In case of monotone missing data the previous expectations reduce to

E-step: monotone missing data
Expectations missing values:

$$Y_{qti} = \begin{cases} Y_{ti}, & t = 1, \ldots, N_i \\ \widehat{\mu}_{qti} + \widehat{\alpha}'_{qi}\widehat{\varepsilon}_{qt(i-1)}, & t = N_i + 1, \ldots, N. \end{cases}$$

Expectations inner products missing values:

$$E\{Y_{ti}Y'_{tj}|Y_{obs}, \widehat{\beta}_q, \widehat{\Sigma}_q\} = Y_{qti}\,(Y_{qtj})' + c_{qtij}, \text{ where}$$

$$c_{qtij} = \begin{cases} 0, & t = 1, \ldots, max(N_i, N_j) \\ \widehat{\Sigma}_{q,ij\cdot obs_t}, & t = max(N_i, N_j) + 1, \ldots, N. \end{cases}$$

5.4.4 M-step

In the maximization step of an EM-algorithm, the loglikelihood of the expected values of all the variables (observed and missing), *i.e.* the completed likelihood, is maximized w.r.t. the parameters characterizing the likelihood. In case of complete observations and identical explanatory variables for all dependent variables, ML estimation and OLS coincide (see Van der Genugten (1988) p. 495, *e.g.*). The maximization step in iteration $q+1$ reads

M-step

$$\begin{aligned}
\widehat{\beta}_{q+1} &= E\{(X'X)^{-1}X'Y|Y_{obs},\widehat{\beta}_q,\widehat{\Sigma}_q\} = (X'X)^{-1}X'Y_q, \\
\widehat{\Sigma}_{q+1} &= E\{\left(Y-X\widehat{\beta}_{q+1}\right)'\left(Y-X\widehat{\beta}_{q+1}\right)/N|Y_{obs},\widehat{\beta}_q,\widehat{\Sigma}_q\} \\
&= \frac{E\{Y'Y|Y_{obs},\beta_q,\Sigma_q\}-\widehat{\beta}'_{q+1}X'X\widehat{\beta}_{q+1}}{N}.
\end{aligned}$$

Since the observations for the first group of dependent variables are complete, it is clear that the MLE's for this group will be obtained after one iteration. It is not clear how many iterations are required for the numerical convergence of the estimates for the other groups. The rate of convergence depends on several factors such as the fraction of missing observations (see McLachan and Krishnan (1997) *e.g.*).

Numerical illustration
We applied the described EM-algorithm to the numerical example of Section 4.2. As starting value we took the ML estimate based solely on the N_r complete observations. The EM-algorithm needed 10 (20) iterations to produce the maximum likelihood estimates, accurate up to two (four) decimals. This is considerably more than the iterative EGLS procedure of Section 5.3.

5.5 One-way MANOVA

5.5.1 The model

We look at the model for one-way MANOVA with factor A having a (≥ 2) levels $A_1, \ldots, A_a$, with n_{ij} ($n_{ij} \geq 1$) observations of the i^{th} group of dependent variables on the j^{th} level. The t^{th} observation on level j of the dependent variables in group i is denoted by Y_{ijt}. If there are r groups, the regression equations read as follow:

$$Y_{ijt} = \mu_{ijt} + \varepsilon_{ijt}, \quad i = 1, \ldots, r, \quad j = 1, \ldots, a, \quad t = 1, \ldots, n_{ij} \tag{5.5.1}$$

where

$$\mu_{ijt} = \mu_{ij} = \beta_{ic} + \beta_{ij}. \tag{5.5.2}$$

We want to interpret β_{ic} as the general level of the i^{th} group of dependent variables and β_{ij} as the specific contribution of level A_j for the i^{th} group of dependent variables. One of the following identifiability conditions is often imposed:

$$\text{(unweighted)} \quad \sum_{j=1}^{a} \beta_{ij} = 0, \text{ for } i = 1, \ldots, r, \tag{5.5.3}$$

$$\text{(weighted)} \quad \sum_{j=1}^{a} n_{ij}\beta_{ij} = 0, \text{ for } i = 1, \ldots, r. \tag{5.5.4}$$

By introducing a dummy variable for each level A_i of A, (5.5.1) and (5.5.2) can be written as a linear regression model. Let

$$X_{jt}^{A} = \begin{cases} 1 & \text{if observation } t \text{ is performed at level } A_j \\ 0 & \text{else.} \end{cases}$$

We will denote the observations of the dummy variables for level j by the vector $X_j^A = [X_{jt}^A]$, and for all levels by $X_A = [X_1^A \ \ldots \ X_a^A]$. Similarly, the matrix with the observations of all the explanatory variables (*i.e.* the constant and the dummy variables) is denoted by $X = [1_N \ X_A]$.

The model assumptions concerning the error terms are those of Chapter 4 (see (4.2.2)). A monotone missing data structure is assumed, so $n_{ij} \geq n_{i+1,j}$. Note that $N_i = \sum_{j=1}^{a} n_{ij}$.

5.5.2 Notation for averages and covariances

In the remainder of the section we will see that the EGLS estimators and their inner products in the MANOVA-tables can be expressed in terms of sample averages and (co)variances. Therefore we introduce symbols to denote these frequently used sample statistics. We denote the sample means by

$$\begin{aligned}
\overline{Y}_{ij\cdot} &= \frac{1}{n_{ij}}\sum_{t=1}^{n_{ij}} Y_{ijt} \quad (\in \mathbb{R}^{m_i\times 1}),\\
\overline{Y}_{iv} &= [\overline{Y}_{i1\cdot} \quad \dots \quad \overline{Y}_{ia\cdot}] \quad (\in \mathbb{R}^{m_i\times a}),\\
\text{(unweighted)}\quad \widehat{Y}_{i\cdot\cdot} &= \frac{1}{a}\sum_{j=1}^{a} \overline{Y}_{ij\cdot} \quad (\in \mathbb{R}^{m_i\times 1}),\\
\text{(weighted)}\quad \overline{Y}_{i\cdot\cdot} &= \frac{1}{N_i}\sum_{j=1}^{a}\sum_{t=1}^{n_{ij}} Y_{ijt} = \sum_{j=1}^{a}\frac{n_{ij}}{N_i}\overline{Y}_{ij\cdot} \quad (\in \mathbb{R}^{m_i\times 1}).
\end{aligned}$$

A similar notation is used for the sample means of the residuals.

We denote the sample (co)variances by

$$\begin{aligned}
\overline{\Sigma}_{(i-1)i} &= \frac{1}{N_i}\widehat{\varepsilon}'_{(i-1)}Y_i - \sum_{j=1}^{a}\frac{n_{ij}}{N_i}\overline{\varepsilon}_{(i-1)j\cdot}\overline{Y}'_{ij\cdot},\\
\overline{\Sigma}_{(i-1)(i-1)} &= \frac{1}{N_i}\widehat{\varepsilon}'_{(i-1)}\widehat{\varepsilon}_{(i-1)} - \sum_{j=1}^{a}\frac{n_{ij}}{N_i}\overline{\varepsilon}_{(i-1)j\cdot}\overline{\varepsilon}'_{(i-1)j\cdot}.
\end{aligned}$$

5.5.3 EGLS estimation

Since EGLS estimation for the first group coincides with OLS estimation, the EGLS estimators for this group are the usual one-way MANOVA-estimators. Regardless of the specific identifiability constraint for the regression coefficients, the OLS projections are

$$\widehat{\mu}_{1jt} = \overline{Y}_{1j\cdot} \text{ and } \widehat{\varepsilon}_{1jt} = Y_{1jt} - \overline{Y}_{1j\cdot}, \quad j = 1,\dots,a, \quad t = 1,\dots,n_{1j}.$$

The unweighted constraint (5.5.3) leads to the following OLS estimators for the regression coefficients

$$\begin{aligned}\widehat{\beta}_{1c} &= \widehat{Y}_{1\cdot\cdot} \quad , \\ \widehat{\beta}_{1j} &= \overline{Y}_{1j\cdot} - \widehat{Y}_{1\cdot\cdot} \quad , \quad j = 1, \ldots, a,\end{aligned}$$

and the weighted constraint (5.5.4) to

$$\begin{aligned}\widehat{\beta}_{1c} &= \overline{Y}_{1\cdot\cdot} \quad , \\ \widehat{\beta}_{1j} &= \overline{Y}_{1j\cdot} - \overline{Y}_{1\cdot\cdot} \quad , \quad j = 1, \ldots, a.\end{aligned}$$

For EGLS estimation for group i $(= 2, \ldots, r)$ the regression equations read

$$Y_i = X_i\beta_i + \widehat{\varepsilon}_{(i-1)}\alpha_i + \varepsilon_i.$$

Since either constraint (5.5.3) or (5.5.4) holds, and $L_{0i} = \mathcal{R}(X_i(\mathcal{N}(C))) = \mathcal{R}(X_{Ai}) = \mathcal{R}(X_i) = L_i$ for both, we can omit the constant term when calculating the EGLS estimators for ν_i and μ_{ij}.

The EGLS estimator for ν_i can easily be determined by means of partial regression. First we regress $[Y_i \quad \widehat{\varepsilon}_{(i-1)}]$ onto X_{Ai}. Since the columns of X_{Ai} are orthogonal, this is straightforward and leads to the centered residuals

$$[Y_i - X_{Ai}\overline{Y}'_{iv\cdot} \quad \widehat{\varepsilon}_{(i-1)} - X_{Ai}\overline{\varepsilon}'_{(i-1)v\cdot}].$$

The second step consists of the regression of these residuals of Y_i onto the corresponding residuals of $\widehat{\varepsilon}_{(i-1)}$. This leads to the (final) residuals of Y_i

$$\begin{aligned}\widehat{\eta}_i &= Y_i - X_{Ai}\overline{Y}'_{iv\cdot} - (\widehat{\varepsilon}_{(i-1)} - X_{Ai}\overline{\varepsilon}'_{(i-1)v\cdot}) \cdot \\ &\quad ((\widehat{\varepsilon}_{(i-1)} - X_{Ai}\overline{\varepsilon}'_{(i-1)v\cdot})'(\widehat{\varepsilon}_{(i-1)} - X_{Ai}\overline{\varepsilon}'_{(i-1)v\cdot}))^{-1} \\ &\quad (\widehat{\varepsilon}_{(i-1)} - X_{Ai}\overline{\varepsilon}'_{(i-1)v\cdot})'(Y_i - X_{Ai}\overline{Y}'_{iv\cdot}) \\ &= Y_i - X_{Ai}\overline{Y}'_{iv\cdot} - (\widehat{\varepsilon}_{(i-1)} - X_{Ai}\overline{\varepsilon}'_{(i-1)v\cdot}) \cdot \\ &\quad (\widehat{\varepsilon}'_{(i-1)}\widehat{\varepsilon}_{(i-1)} - \sum_{j=1}^{a} n_{ij}\overline{\varepsilon}_{(i-1)j\cdot}\overline{\varepsilon}'_{(i-1)j\cdot})^{-1}(\widehat{\varepsilon}'_{(i-1)}Y_i - \sum_{j=1}^{a} n_{ij}\overline{\varepsilon}_{(i-1)j\cdot}\overline{Y}'_{ij\cdot}) \\ &= Y_i - X_{Ai}\overline{Y}'_{iv\cdot} - (\widehat{\varepsilon}_{(i-1)} - X_{Ai}\overline{\varepsilon}'_{(i-1)v\cdot})\overline{\Sigma}^{-1}_{(i-1)(i-1)}\overline{\Sigma}_{(i-1)i}.\end{aligned}$$

Since $Y_i = \widehat{\nu}_i + \widehat{\eta}_i$, this leads to

$$\widehat{\nu}_i = X_{Ai}\overline{Y}_{iv\cdot} + (\widehat{\varepsilon}_{(i-1)} - X_{Ai}\overline{\varepsilon}'_{(i-1)v\cdot})\overline{\Sigma}^{-1}_{(i-1)(i-1)}\overline{\Sigma}_{(i-1)i}, \tag{5.5.5}$$

or equivalently

$$\widehat{\nu}_{ijt} = \overline{Y}'_{ij\cdot} - \overline{\Sigma}_{i(i-1)}\overline{\Sigma}^{-1}_{(i-1)(i-1)}\left(\overline{\varepsilon}_{(i-1)j\cdot} - \widehat{\varepsilon}_{(i-1)jt}\right).$$

This expression and relation (4.3.12) lead to the estimator of the mean of group i for level A_j

$$\widehat{\mu}_{ij} \;=\; \overline{Y}_{ij\cdot} - \overline{\Sigma}_{i(i-1)}\overline{\Sigma}^{-1}_{(i-1)(i-1)}\overline{\varepsilon}_{(i-1)j\cdot} \quad .$$

The EGLS estimator $\widehat{\mu}_{ij}$ and the constraint (5.5.3) or (5.5.4) give the EGLS estimators for the regression coefficients. In case of constraint (5.5.3) the EGLS estimators for the regression coefficients are

$$\begin{aligned}
\widehat{\beta}_{ic} &= \widehat{Y}_{i\cdot\cdot} - \overline{\Sigma}_{i(i-1)}\overline{\Sigma}^{-1}_{(i-1)(i-1)}\widehat{\varepsilon}_{(i-1)\cdot\cdot} \quad , \\
\widehat{\beta}_{ij} &= \overline{Y}_{ij\cdot} - \widehat{Y}_{i\cdot\cdot} - \overline{\Sigma}_{i(i-1)}\overline{\Sigma}^{-1}_{(i-1)(i-1)}(\overline{\varepsilon}_{(i-1)j\cdot} - \widehat{\varepsilon}_{(i-1)\cdot\cdot}) \, ,
\end{aligned}$$

and in case of constraint (5.5.4)

$$\begin{aligned}
\widehat{\beta}_{ic} &= \overline{Y}_{i\cdot\cdot} - \overline{\Sigma}_{i(i-1)}\overline{\Sigma}^{-1}_{(i-1)(i-1)}\overline{\varepsilon}_{(i-1)\cdot\cdot} \\
\widehat{\beta}_{ij} &= \overline{Y}_{ij\cdot} - \overline{Y}_{i\cdot\cdot} - \overline{\Sigma}_{i(i-1)}\overline{\Sigma}^{-1}_{(i-1)(i-1)}(\overline{\varepsilon}_{(i-1)j\cdot} - \overline{\varepsilon}_{(i-1)\cdot\cdot}).
\end{aligned}$$

The EGLS estimators $\widehat{\mu}_i$ and $\widehat{\beta}$ are the usual one-way MANOVA-estimators plus a deviation. In case of complete data, $\overline{\varepsilon}_{(i-1)j\cdot} = 0$ for all j and thus $\overline{\varepsilon}_{(i-1)\cdot\cdot} = 0$ and $\widehat{\varepsilon}_{(i-1)\cdot\cdot} = 0$. As a consequence, $\widehat{\mu}_{ij}$ and $\widehat{\beta}$ reduce to the 'regular' one-way MANOVA-estimators.

If some observations are missing but not for level A_j ($n_{ij} = n_{1j}$), then $\overline{\varepsilon}_{(i-1)j\cdot} = 0$ but $\widehat{\varepsilon}_{(i-1)\cdot\cdot} \neq 0$ and $\overline{\varepsilon}_{(i-1)\cdot\cdot} \neq 0$. Hence $\widehat{\mu}_{ij} = \overline{Y}_{ij\cdot}$ but $\widehat{\beta}$ does not reduce to the 'regular' one-way MANOVA-estimator.

5.5.4 MANOVA-tables

In Section 4.6 we looked at the collection of MANOVA-tables for (general) multivariate regression with consecutively added dependent variables. These MANOVA-tables (see for example Table 4.6.1) contain the inner products of the unconstrained and constrained projections and the corresponding degrees of freedom. In this section, we only present the MANOVA-tables for the model test (*i.e.* the null hypothesis assumes all regression coefficients to be zero except the constant term). Table 5.5.1 contains the relevant information for the model test.

Model	**Space**	**SS**	**DF**
C. model	$\widetilde{L}_{(i)}$	$\widetilde{\nu}_i'\widetilde{\nu}_i$	$a-1$
Error	$L_{(i)}^{\perp}$	$\widehat{\eta}_i'\widehat{\eta}_i$	N_i-a
C. total	$\mathcal{R}(1_{N_i})^{\perp}$	$\widetilde{Y}_i'\widetilde{Y}_i$	N_i-1
Mean	$\mathcal{R}(1_{N_i})$	$N_i\overline{Y}_{i\cdot\cdot}'\overline{Y}_{i\cdot\cdot}$	1
Total	$I\!R^{N_i}$	$Y_i'Y_i$	N_i

Table 5.5.1: Collection of centered MANOVA-tables ($i=2,\ldots,r$)

To determine the exact expressions for the inner products of the MANOVA-table, we first determine $\widehat{\nu}_i'\widehat{\nu}_i$. Since $\widehat{\nu}_i$ in (5.5.5) is the sum of two orthogonal terms, its inner product is the sum of the two corresponding inner products:

$$\begin{aligned}\widehat{\nu}_i'\widehat{\nu}_i &= (X_{Ai}\overline{Y}_{iv\cdot}')'(X_{Ai}\overline{Y}_{iv\cdot}')+\overline{\Sigma}_{i(i-1)}\overline{\Sigma}_{(i-1)(i-1)}^{-1}(\widehat{\varepsilon}_{(i-1)}-X_{Ai}\overline{\varepsilon}_{(i-1)v\cdot})'\cdot\\ &\quad (\widehat{\varepsilon}_{(i-1)}-X_{Ai}\overline{\varepsilon}_{(i-1)v\cdot}')\overline{\Sigma}_{(i-1)(i-1)}^{-1}\overline{\Sigma}_{(i-1)i}\\ &= \sum_{j=1}^{a} n_{ij}\overline{Y}_{ij\cdot}\overline{Y}_{ij\cdot}'+N_i\overline{\Sigma}_{i(i-1)}\overline{\Sigma}_{(i-1)(i-1)}^{-1}\overline{\Sigma}_{(i-1)i}.\end{aligned}$$

The inner products of the EGLS residuals are

$$\begin{aligned}\widehat{\eta}_i'\widehat{\eta}_i &= Y_i'Y_i-\widehat{\nu}_i'\widehat{\nu}_i\\ &= \sum_{j=1}^{a}\sum_{t=1}^{n_{ij}}(Y_{ijt}-\overline{Y}_{ij\cdot})(Y_{ijt}-\overline{Y}_{ij\cdot})'-N_i\overline{\Sigma}_{i(i-1)}\overline{\Sigma}_{(i-1)(i-1)}^{-1}\overline{\Sigma}_{(i-1)i}.\end{aligned}$$

Since the EGLS residuals are already centered, the centered inner products $\widetilde{\nu}_i'\widetilde{\nu}_i$ of the MANOVA-table can be determined as the difference between $\widetilde{Y}_i'\widetilde{Y}_i$ and $\widehat{\eta}_i'\widehat{\eta}_i$. The inner products of the centered dependent variables are

$$\widetilde{Y}_i'\widetilde{Y}_i = \sum_{j=1}^{a}\sum_{t=1}^{n_{ij}}(Y_{ijt} - \overline{Y}_{i\cdot\cdot})(Y_{ijt} - \overline{Y}_{i\cdot\cdot})',$$

which leads to

$$\widetilde{\nu}_i'\widetilde{\nu}_i \;\; = \;\; \sum_{j=1}^{a} n_{ij}(\overline{Y}_{ij\cdot} - \overline{Y}_{i\cdot\cdot})(\overline{Y}_{ij\cdot} - \overline{Y}_{i\cdot\cdot})' + N_i\overline{\Sigma}_{i(i-1)}\overline{\Sigma}_{(i-1)(i-1)}^{-1}\overline{\Sigma}_{(i-1)i}.$$

The first terms of $\widetilde{\nu}_i'\widetilde{\nu}_i$ and $\widehat{\eta}_i'\widehat{\eta}_i$ are the inner products between the samples and within the samples, respectively.

5.6 Conclusions

This chapter discussed several features of the model for multivariate regression with consecutively added dependent variables. We proved that all estimators of the previous chapter, and new classes of covariance estimators are consistent. In Section 4.4 we investigated the relative efficiency of the estimators for the regression coefficients, but we have not studied the (asymptotic) relative efficiency of the estimators for the (co)variances yet. From general theory it is known that the MLE's are asymptotic efficient. Are there also asymptotic efficient estimators in the new classes of covariance estimators? If not, which estimators in the new classes are the best in terms of efficiency? We leave these questions for further research.

We also described two alternative, often used, estimation techniques. Although these procedures numerically converge to the ML estimates, they do not result in closed form estimators for the coefficients. Therefore, our estimation technique of Section 4.3 is simpler, more straightforward, and much faster.

Finally, we also looked at a special case of the model of Chapter 4: one-way MANOVA. This simple generalization of the model with only the constant term as explanatory variable, resulted in quite complicated expressions for the EGLS estimators.

Chapter 6

Mixed models

6.1 Introduction

The previous chapters discussed models (with applications to repeated audit controls) with either categorical or continuous variables. However, in audit practice the records are often correct (*i.e.* the error is zero); but if they are incorrect, the errors can take many different values (see Johnson *et al*. (1981) or Neter *et al*. (1985) *e.g.* for a more detailed discussion). The resulting error hence has a mixed distribution; we therefore will call models for this frequently occurring situation mixed models.

The model with continuous errors and a probability mass in zero has been discussed in literature. Cox and Snell (1979) derived Bayesian estimators and upper limits for a model with non-negative errors and a probability mass in zero. Moors (1983) and Moors and Janssens (1989) expanded on this. Estimators for continuous, but not necessarily positive, errors with a point mass in zero were derived by Fienberg *et al*. (1977), Tamura and Frost (1986), Tamura (1988) and Laws and O'Hagan (2000). However, they all assume one audit round with an infallible auditor. This in contrast to Barnett *et al*. (2001) who discussed a repeated audit control with two rounds. First a model for the classification frequencies was presented and MLE's for the classification probabilities were derived. Further, based on the observed errors, several estimators for the mean value of the errors in the population were proposed; no relation was specified between the size of the

non-zero errors and the (registered) values of the records.

Section 6.2 introduces our mixed model for a repeated audit control with two rounds. In Section 6.2.2 the model of Chapter 2 for the classification probabilities is extended slightly; the resulting model is identical to the model of Barnett *et al*. (2001). Conditional on the classification of a record, we specify regression models for the non-zero error in Section 6.2.3. These conditional linear regression models are similar to the one of Chapter 4.

In Section 6.3 the estimation techniques of Chapters 2 and 4 are used to determine estimators for the classification probabilities and regression parameters, respectively. The OLS estimators and MLE's for the parameters of the conditional regression models are compared by means of simulation. Section 6.4 discusses estimators for the mean value of the errors in the population. We present the MLE for our model and briefly discuss the estimators of Barnett *et al*. (2001). All the estimators are compared by means of simulation. The final Section 6.5 contains our main conclusions and ideas for further research.

6.2 The model

6.2.1 Notation

Define the random variable A_0 as the registered value (or the so called book value) of a random record. The random variables A_1 and A_2 are defined as the values of a random record according to the first auditor and the expert, respectively. Since the expert is assumed to be infallible A_2 is the true value. We denote the book and audit values of record t by $A_{t0},\ A_{t1}$, and A_{t2}, respectively.

As in Chapter 2 the first auditor checks the records of a random sample (drawn with replacement) of predetermined size n_1; a subsample of (possible random) size $N_2 \leq n_1$ is checked again by the expert. Now the values of $(A_{t0},\ A_{t1},\ A_{t2})$ are available for the N_2 double checked sample records, while for the $n_1 - N_2$ single checked sample records only $(A_{t0},\ A_{t1})$ are available. Since in practice the book values are known for all records of the population, we will assume that A_{t0} is known for the whole population.

In Section 3.2.3 we discussed two different approaches to determine N_2: random and stratified sampling. Both methods can be applied in this chapter. However, we will not elaborate on this difference since the sampling method does not influence the MLE's (see Theorem 3.3.2).

Our model is constructed from an absolute model for the classification probabilities and a conditional model for the audit values. First all records are classified into five groups, based on the question whether the two audit values and the book value are identical. In Section 6.2.2 we give our model for the corresponding classification probabilities. If all three values coincide, no further steps are necessary. In the four other cases, we still need to specify models for one of the audit values, or both. Section 6.2.3 describes these conditional regression models.

6.2.2 Classifications

As in Chapter 2, π_0 (π_1) is the probability that the auditor classifies a random record as 'incorrect' ('correct'). With conditional probability $\pi_{0|0}$ ($\pi_{1|1}$) the 'incorrect' ('correct') record is indeed incorrect (correct). With conditional probability $\pi_{1|0}$ ($\pi_{0|1}$) the 'incorrect' ('correct') record was misclassified by the auditor and is correct (incorrect) after all. Joint probabilities as $\pi_{01} = \pi_0\pi_{1|0}$ (a random record being classified as 'incorrect' by the auditor and as correct by the expert) follow from these; compare Figure 2.2.1.

So far our model for the classification probabilities is identical to the model of Chapter 2. However, now we are interested not only in the fraction errors but also in the size of the errors; an additional subdivision is therefore necessary. If the auditor correctly concludes that a record is in error, two possibilities remain: (s)he is correct about the size of the error, or not. Accordingly, we introduce the probabilities $\pi_{0e|0}$ ($\pi_{0u|0}$) for the events that the error size indicated by the auditor is *e*qual (*u*nequal) to the true error. So $\pi_{0|0} = \pi_{0e|0} + \pi_{0u|0}$ and $\pi_{00} = \pi_{00e} + \pi_{00u}$.

The foregoing classifications and probabilities can be expressed in terms of book and audit values. For example

$$\pi_{0u|0} = Pr(A_0 \neq A_2, A_1 \neq A_2 | A_0 \neq A_1).$$

Table 6.2.1 gives an overview of the five possible classifications and their probabilities.

Classification	Probability
1. $A_0 = A_1$, $A_0 = A_2$	π_{11}
2. $A_0 = A_1$, $A_0 \neq A_2$	π_{10}
3. $A_0 \neq A_1$, $A_0 = A_2$	π_{01}
4. $A_0 \neq A_1$, $A_0 \neq A_2$, $A_1 = A_2$	π_{00e}
5. $A_0 \neq A_1$, $A_0 \neq A_2$, $A_1 \neq A_2$	π_{00u}

Table 6.2.1: Classifications and probabilities

As in Chapter 2, we denote the sample classification frequencies by the symbol C with the same subindices as the corresponding probabilities π (see Table 6.2.1). Figure 6.2.1 gives an overview of the sample frequencies and probabilities (compare Figure 2.2.1).

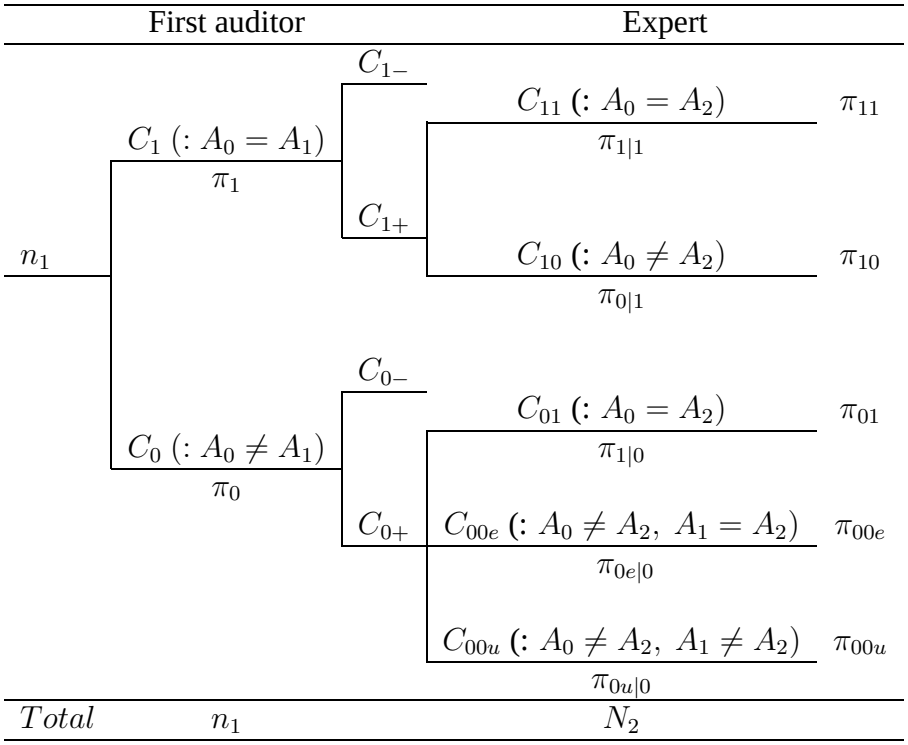


Figure 6.2.1: Classification frequencies and probabilities

6.2.3 Conditional regression

Since the book value is available for each record, it is only necessary to specify a conditional model for A_{t1} given $A_{t1} \neq A_{t0}$. Whether this is the case follows from the classification of record t. If the book and audit value do not coincide, it seems reasonable to assume that the book value influences the audit value. So we assume

$$A_{t1} = \beta_0' \begin{bmatrix} 1 \\ A_{t0} \end{bmatrix} + \varepsilon_t, \text{ with } E(\varepsilon_t | A_{t0}) = 0 \text{ given } A_{t0} \neq A_{t1},$$

for some (regression) coefficient β_0. Here we omit in our notation for the expectation (and in the following for the variance) the condition $A_{t0} \neq A_{t1}$. Moreover, we assume a constant variance ($V(\varepsilon_t | A_{t0}) = \sigma_0^2$) and no correlation between records.

We only need to specify a model for A_{t2} if the true value does not coincide with the book or previous audit value. This is the case for the classifications 2 and 5 in Table 6.2.1. For both classifications we assume linear regression models, which are not necessary identical: after all, the first auditor missing an error might indicate that the error is quite small, while the first auditor finding an error (but not the true one) might indicate a large or complicated error. We assume

$$A_{t2} = \beta_1' \begin{bmatrix} 1 \\ A_{t0} \end{bmatrix} + \varepsilon_t, \text{ with } E(\varepsilon_t | A_{t0}) = 0 \text{ given } \begin{cases} A_{t0} = A_{t1} \\ A_{t0} \neq A_{t2} \end{cases},$$

for some (regression) coefficient β_1. Again we assume that the variance of the error terms is constant ($V(\varepsilon_t | A_{t0}) = \sigma_1^2$) and that there is no correlation between records.

Similarly, we assume

$$A_{t2} = \beta_{0u}' \begin{bmatrix} 1 \\ A_{t0} \end{bmatrix} + \varepsilon_t, \text{ with } E(\varepsilon_t | A_{t0}) = 0 \text{ given } \begin{cases} A_{t0} \neq A_{t1} \\ A_{t0} \neq A_{t2} \\ A_{t1} \neq A_{t2} \end{cases},$$

for some (regression) coefficient β_{0u}. Although we assume again a constant variance ($V(\varepsilon_t | A_{t0}) = \sigma_{0u}^2$) and no correlation between different records, we do not impose restrictions on the correlation between the audit and true value per record (or equivalently, the covariance σ_{12}).

Table 6.2.2 gives an overview of the explanatory and dependent variables of the conditional regression models in the notation of Chapter 4.

Parameters	β_1, σ_1^2	β_0, σ_0^2	β_{0u}, σ_{0u}^2, σ_{12}
dependent variables Y_{ti}	A_{t2}	A_{t1}	A_{t2}
explanatory variables $[X_{t1}\ X_{t2}]$	$[1\ A_{t0}]$	$[1\ A_{t0}]$	$[1\ A_{t0}]$
previous error terms $\varepsilon_{t(i-1)}$	-	-	$A_{t1} - \beta_0' \begin{bmatrix} 1 \\ A_{t0} \end{bmatrix}$
number of observations N_i	C_{10}	C_0	C_{00u}

Table 6.2.2: Explanatory and dependent variables

In all our conditional regression models, the explanatory variables consist of the constant and the book value. The conditional model given $A_{t0} = A_{t1}$, has the true value as dependent variable. The other two conditional models (given $A_{t0} \neq A_{t1}$) form a bivariate regression model with monotone missing observations: for the first dependent variable (the value according to the first auditor) C_0 observations are available, while for the second dependent variable (the true value) only C_{00u} observations are available.

We will use the estimation techniques of Chapter 4 to determine estimators for the parameters of the conditional regression models.

Table 6.2.3 gives an overview of the conditional regression models for all classifications. This overview will be especially useful for the estimation of the mean true value in Section 6.4.

Classification	**Conditional regression model**
$A_0 = A_1$,$A_0 = A_2$	-
$A_0 = A_1$, $A_0 \neq A_2$	$A_{t2} = \beta_1' \begin{bmatrix} 1 \\ A_{t0} \end{bmatrix} + \varepsilon_t$, $E(\varepsilon_t \vert A_{t0}) = 0$, $Cov(\varepsilon_t \vert A_{t0}) = \sigma_1^2$,
$A_0 \neq A_1$, $A_0 = A_2$	$A_{t1} = \beta_0' \begin{bmatrix} 1 \\ A_{t0} \end{bmatrix} + \varepsilon_t$ $E(\varepsilon_t \vert A_{t0}) = 0$, $Cov(\varepsilon_t \vert A_{t0}) = \sigma_0^2$,
$A_0 \neq A_1$, $A_0 \neq A_2$, $A_1 = A_2$	$A_{t1} = \beta_0' \begin{bmatrix} 1 \\ A_{t0} \end{bmatrix} + \varepsilon_t$, $E(\varepsilon_t \vert A_{t0}) = 0$, $Cov(\varepsilon_t \vert A_{t0}) = \sigma_0^2$,
$A_0 \neq A_1$, $A_0 \neq A_2$, $A_1 \neq A_2$	$\begin{bmatrix} A_{t1} \\ A_{t2} \end{bmatrix} = \begin{bmatrix} \beta_0' \\ \beta_{0u}' \end{bmatrix} \begin{bmatrix} 1 \\ A_{t0} \end{bmatrix} + \varepsilon_t$, $E(\varepsilon_t \vert A_{t0}) = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$, $Cov(\varepsilon_t \vert A_{t0}) = \begin{bmatrix} \sigma_0^2 & \sigma_{12} \\ \sigma_{12} & \sigma_{0u}^2 \end{bmatrix}$

Table 6.2.3: Conditional regression models

6.3 Estimation of the model parameters

6.3.1 Classification probabilities

The classification frequencies have binomial and multinomial distributions similar to (2.2.4). So the MLE's for the classification probabilities are the sample fractions (compare (3.3.3)):

$$\begin{cases} \widehat{\Pi}_1 = \dfrac{C_1}{n_1}, \qquad \widehat{\Pi}_0 = \dfrac{C_0}{n_1} \\ \widehat{\Pi}_{1|1} = \dfrac{C_{11}}{C_{1+}}, \quad \widehat{\Pi}_{0|1} = \dfrac{C_{10}}{C_{1+}} \\ \widehat{\Pi}_{1|0} = \dfrac{C_{01}}{C_{0+}}, \quad \widehat{\Pi}_{0e|0} = \dfrac{C_{00e}}{C_{0+}}, \quad \widehat{\Pi}_{0u|0} = \dfrac{C_{00u}}{C_{0+}}. \end{cases} \tag{6.3.1}$$

These MLE's can be found in Barnett *et al.* (2001) as well.

If C_{0+} or C_{1+} is zero, not all MLE's in (6.3.1) are defined. See Section 3.3.3 for a more detailed discussion of this situation and possible solutions.

6.3.2 Regression parameters

The estimators for the regression parameters of the conditional regression models in Section 6.2.3 can be determined by means of the estimation procedures in Section 4.3.2. In terms of general dependent variables Y and explanatory variable X, the OLS estimators for the regression coefficients and (co)variances are (4.3.5) and (4.3.7), respectively; under the normality assumption the MLE 's are (4.3.20) and (4.3.26). Table 6.2.2 gives an overview of the dependent and explanatory variables for the parameters in our conditional regression models. For completeness, we include the OLS estimators and MLE's in terms of the book and audit values in Appendix 6.6.1.

The MLE's for β_1 and β_0 coincide with the OLS estimators. The MLE's for σ_1^2 and σ_0^2 differ from the OLS estimators solely by the denominator: the MLE's are the inner products of the residuals divided by the number of observations, while the OLS estimators are the same inner products divided by the degrees of freedom. Only with respect to β_{0u}, σ_{0u}^2 and σ_{12} the MLE's differ essentially from the OLS estimators. In the next subsection we study the relative efficiency of the OLS estimators and MLE's for these parameters by simulation.

6.3.3 Practical example

As in Chapter 2, the practical example concerns the Dutch social security payments. However, now we consider another case study where also error sizes are observed. The population consists of 587 social security payments with mean 9.0418 and standard deviation 8.5726 (both in 1000's of Dutch guilders). An internal auditor checks all 587 social security payments; an external auditor (the expert) checks a subsample of size 60 once more. We will assume here that the 587 payments checked by the first auditor constitute a sample from a large population. In this context the variable A_0 is the social security payment which actually has been paid, A_1 (A_2) is the social security payment which should have been paid according to the first auditor (expert). Table 6.3.1 contains the classification quantities of the control.

	Total	Single checked sample	Double checked sample		
			Expert		
First auditor			Total	correct	incorrect
'correct'	$c_1 = 551$	$c_{1-} = 493$	$c_{1+} = 58$	$c_{11} = 55$	$c_{10} = 3$
'incorrect'	$c_0 = 36$	$c_{0-} = 34$	$c_{0+} = 2$	$c_{01} = 0$	$c_{00e} = 2$
Total	$n_1 = 587$	$n_1 - n_2 = 527$	$n_2 = 60$	$c_{+1} = 55$	$c_{+0} = 5$

Table 6.3.1: CTSV example

In the double checked sample the first auditor did not make up errors, missed three errors and found two (true) errors; the expert confirmed the size of the latter errors.

For these classification frequencies, (6.3.1) results in the ML estimates

$$\widehat{\pi}_{11} = 0.8901, \quad \widehat{\pi}_{10} = 0.0486, \quad \widehat{\pi}_{01} = 0, \quad \widehat{\pi}_{00e} = 0.0613, \quad \widehat{\pi}_{00u} = 0.$$

The ML estimates for the regression parameters are determined from the sample observations of A_{t0}, A_{t1} and A_{t2}. Since there are no sample records with $\{A_{t0} \neq A_{t1}, A_{t0} \neq A_{t1}, A_{t1} \neq A_{t2}\}$ (*i.e.* $c_{00u} = 0$), the parameters β_{0u}, σ^2_{0u} and σ_{12} can not be estimated. The ML estimates for the other regression parameters are

$$\widehat{\beta}_1 = \begin{bmatrix} -14.7107 \\ -0.8275 \end{bmatrix}, \quad \widehat{\sigma}^2_1 = 53.5911, \quad \widehat{\beta}_0 = \begin{bmatrix} -0.6807 \\ 0.8808 \end{bmatrix}, \quad \widehat{\sigma}^2_0 = 17.3533.$$

These ML estimates are used in our simulations to study the relative efficiency of the OLS estimators and MLE's for β_{0u}, σ^2_{0u} and σ_{12}.

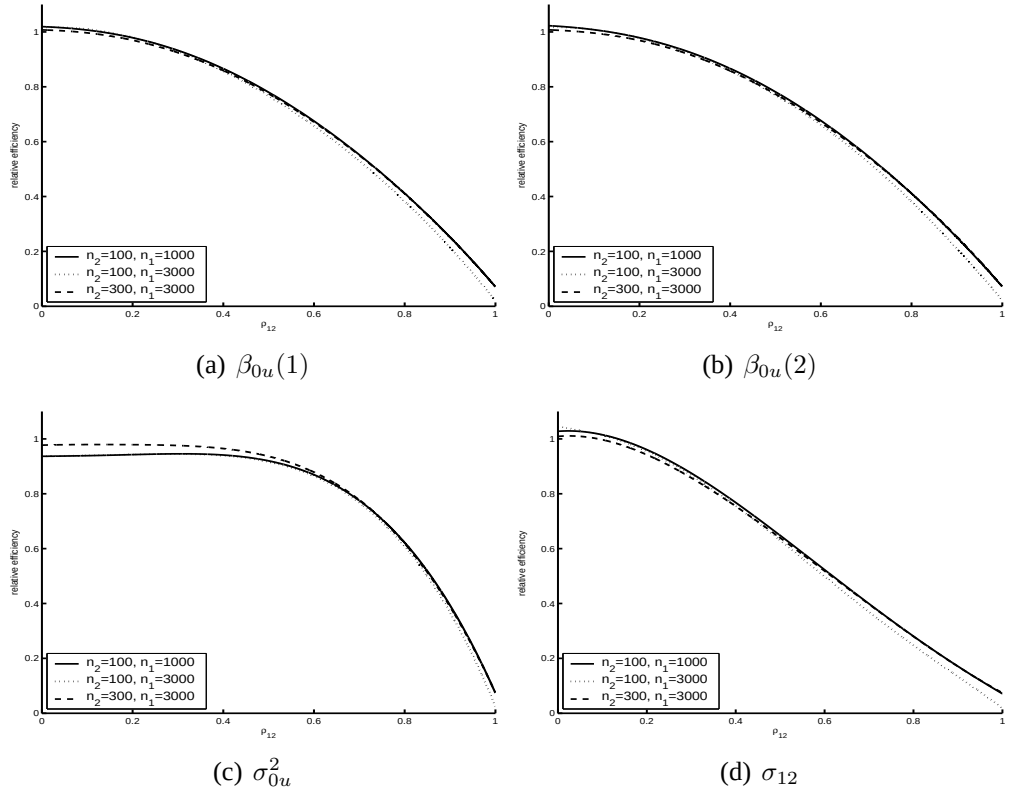
The difference between OLS and ML estimation mainly stems from the treatment of the C_{00u} observations where the auditor correctly identifies an error, but errs in its size. Hence in the simulation study, we use a value of the classification probability π_{00u} which is unlikely to lead to zero observations in this category:

$$\pi_{11} = \pi_{10} = \pi_{01} = \pi_{00e} = 0.1, \quad \pi_{00u} = 0.6.$$

We take the regression parameters equal to the corresponding ML estimates of the practical example; in addition we assume that β_{0u} (σ^2_{0u}) is equal to β_0 (σ^2_0). Since we expect the correlation between A_{t1} and A_{t2} (given $\{A_{t0} \neq A_{t1}, A_{t0} \neq A_{t1}, A_{t1} \neq A_{t2}\}$) to be important for the relative efficiency, we look at different

values for the correlation coefficient (ρ_{12}); this determines as well the covariance $\sigma_{12} = \rho_{12}\sigma_0\sigma_{0u}$.

We simulate the book values from a normal distribution with mean 9.0418 and standard deviation 8.5726 from the practical example. The audit values are also drawn from (multi)normal distributions. To determine the effect of the sample sizes, we have simulated data (each with runsize 10,000) for three different situations: (a) $n_2 = 100, n_1 = 1000$, (b) $n_2 = 100, n_1 = 3000$ and (c) $n_2 = 300, n_1 = 3000$. Figure 6.3.1 contains the smoothed curves of the relative efficiency for the different parameters as function of ρ_{12}. Note that each graph contains three curves, which however often partly coincide.



(a) $\beta_{0u}(1)$ (b) $\beta_{0u}(2)$

(c) σ^2_{0u} (d) σ_{12}

Figure 6.3.1: Relative efficiency of OLS in relation to ML

The first and second graph show the relative efficiency for the first and second component of β_{0u}, respectively. These graphs show the same pattern as Figure

4.4.1 and hence confirm our findings of Section 4.4. For low values of the correlation coefficient, there is hardly any difference in efficiency between the two estimators; for high values, $\widehat{\beta}_{0u}$ is much more efficient than b_{0u}. This difference in efficiency increases with the missing data ratio. Note that the difference seems not to depend on the absolute sample sizes themselves, only on this ratio $1 - n_2/n_1$.

The third and fourth graph, for σ^2_{0u} and σ_{12}, show a similar picture as the first two. This is understandable since the MLE's $\widehat{\sigma}^2_{0u}$ and $\widehat{\sigma}_{12}$ are functions of $\widehat{\sigma}^2_0$ which is based on all n_1 observations.

6.4 Estimation of the mean true value

6.4.1 Notation

In a repeated audit control, the main parameter of interest is often the mean true value in the population or equivalently the total true value in the population. The mean population error size is the difference between the mean population book value μ_0 and the mean population true value, μ_2: $\mu_0 - \mu_2$. Since we assume that the book values are available for all population elements, the estimator for the mean error size is obtained by subtracting the estimator for μ_2 from the known parameter μ_0.

In Section 6.4.2 we propose an estimator for μ_2 based on our model. Section 6.4.3 discusses several estimators of Barnett *et al.* (2001). All four estimators are compared by simulation in Section 6.4.4.

We use the following notation for sample averages and regression coefficients

$$\begin{aligned} \overline{A}_g^{(C_{ij})} &= \frac{1}{C_{ij}} \sum^{C_{ij}} A_{tg}, \\ \widehat{\alpha}_{gh}^{(C_{ij})} &= \frac{\sum^{C_{ij}} (A_{tg} - \overline{A}_g^{(C_{ij})})(A_{th} - \overline{A}_h^{(C_{ij})})}{\sum^{C_{ij}} \left(A_{tg} - \overline{A}_g^{(C_{ij})} \right)^2}. \end{aligned}$$

The symbol θ will denote all model parameters, *i.e.* all classification probabilities and regression parameters; the MLE for θ is denoted by $\widehat{\theta}$.

6.4.2 A new estimator

A new estimator for μ_2 is the average of the observed and predicted true values of all population elements:

$$\widehat{\mu}_2 = \frac{1}{n_p} \sum_{t=1}^{n_p} \widehat{A}_{t2}, \tag{6.4.1}$$

with

$$\widehat{A}_{t2} = \begin{cases} A_{t2}, & \text{if } t = 1, \ldots, N_2 \\ E\{A_{t2} | A_{t0}, A_{t1}, A_{t0} = A_{t1}, \widehat{\theta}\}, & \text{if } t = N_2 + 1, \ldots, n_1 \text{ and } A_{t0} = A_{t1} \\ E\{A_{t2} | A_{t0}, A_{t1}, A_{t0} \neq A_{t1}, \widehat{\theta}\}, & \text{if } t = N_2 + 1, \ldots, n_1 \text{ and } A_{t0} \neq A_{t1} \\ E\{A_{t2} | A_{t0}, \widehat{\theta}\}, & \text{else.} \end{cases}$$

Each missing A_{t2} is estimated by its conditional expectation (under the normality assumption) given the observations and the (estimated) parameter values. The conditional expectations differ per classification (see Table 6.2.3) and are given in Appendix 6.6.2.

The advantage of this estimator is that it distinguishes the different classifications and it uses all available sample and population information. It also shares some nice properties with the MLE's which have been derived in Chapter 5.

6.4.3 Estimators Barnett

Although Barnett *et al.* (2001) did not specify a model for the size of the errors, several estimators for μ_2 (or $\mu_0 - \mu_2$) were proposed: the regression estimator, the post-stratification estimator and the estimator from non-overlapping samples.

Similar to (6.4.1), the regression estimator for μ_2 is the average of the observed and predicted A_{t2} of all population elements. However, the predictions for the A_{t2} differ from ours. The regression estimator $\widehat{\mu}_{2r}$, used by Barnett *et al.* (2001) equation (17), equals

$$\widehat{\mu}_{2r} = \overline{A}_2^{(N_2)} + (\overline{A}_1^{(N_1)} - \overline{A}_1^{(N_2)})\widehat{\alpha}_{12}^{(N_2)} + (\mu_0 - \overline{A}_0^{(N_1)})\widehat{\alpha}_{01}^{(N_1)}\widehat{\alpha}_{12}^{(N_2)}. \tag{6.4.2}$$

This estimator is quite logical in case of the following model:

$$\begin{bmatrix} A_{t0} \\ A_{t1} \\ A_{t2} \end{bmatrix} = \beta' + \varepsilon_t, \text{ with } E\{\varepsilon_t\} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \quad Var\{\varepsilon_t\} = \begin{bmatrix} \sigma_{00} & \sigma_{01} & 0 \\ \sigma_{01} & \sigma_{11} & \sigma_{12} \\ 0 & \sigma_{12} & \sigma_{22} \end{bmatrix}.$$

Note, however that this model contradicts the model for the classification probabilities, since it does not distinguish the different classifications. This in contrast to the post-stratification estimator for μ_2 (see Barnett *et al.* (2001) equation (21))

$$\widehat{\pi}_{11}\mu_0 + \widehat{\pi}_{10}\overline{A}_2^{(N_2)} + \widehat{\pi}_{01}\mu_0 + \widehat{\pi}_{00e}\overline{A}_1^{(N_1)} + \widehat{\pi}_{00u}\overline{A}_2^{(N_2)}.$$

This estimator is the sum of the MLE's for the classification probabilities times the estimator for the mean true value of elements with that classification. The disadvantage of this estimator is that the estimators for the mean values per classification can be quite biased. Therefore we propose an alternative estimator $\widehat{\mu}_{2p}$ with the same structure but with different estimators for the stratum means

$$\widehat{\mu}_{2p} = \widehat{\pi}_{11}\overline{A}_2^{(C_{11})} + \widehat{\pi}_{10}\overline{A}_2^{(C_{10})} + \widehat{\pi}_{01}\overline{A}_2^{(C_{01})} + \widehat{\pi}_{00e}\overline{A}_2^{(C_{00e})} + \widehat{\pi}_{00u}\overline{A}_2^{(C_{00u})} \quad (6.4.3)$$

(although it is not mentioned explicitly in their paper, this seems to be the estimator which Barnett *et al.* (2001) used in their simulations). The disadvantage of this post-stratification estimator is that it uses the sample information of the single checked elements solely for the estimation of the classification probabilities; the estimation of the stratum means is only based on the double checked sample.

The last estimator $\widehat{\mu}_{2w}$ uses information from both single and double checked sample elements (see Barnett *et al.* (2001) equation (25))

$$\begin{aligned} \widehat{\mu}_{2w} &= \mu_0 - \frac{N_2}{n_1}(\overline{A}_0^{(N_2)} - \overline{A}_2^{(N_2)}) \\ &\quad - \frac{n_1 - N_2}{n_1}\frac{C_{0-}\widehat{\pi}_{0|0} + C_{1-}\widehat{\pi}_{0|1}}{C_{0-}}(\overline{A}_0^{(N_1-N_2)} - \overline{A}_1^{(N_1-N_2)}). \end{aligned} \quad (6.4.4)$$

This estimator is μ_0 minus the weighted average of the mean error size of the double checked elements and, the mean error size of the single checked sample elements according to the auditor multiplied by a correction factor for the misclassifications. Theorem 6.4.1 shows that $\widehat{\mu}_{2w}$ is not always consistent.

Theorem 6.4.1. *In case of random sampling* $\widehat{\mu}_{2w} \xrightarrow{P} \mu_2$ *if and only if* $E\{A_{t0} - A_{t1}|A_{t0} \neq A_{t1}\} = E\{A_{t0} - A_{t2}|A_{t0} \neq A_{t2}\}$.

Proof. As in Chapters 2 and 3 we denote the fraction incorrect elements in the population by $p_0 (= \pi_{10} + \pi_{00})$.

Since sample means converge to their expectations in case of random sampling, it follows that

$$\overline{A}_0^{(N_2)} - \overline{A}_2^{(N_2)} \xrightarrow{P} \mu_0 - \mu_2, \qquad \overline{A}_0^{(N_1-N_2)} - \overline{A}_1^{(N_1-N_2)} \xrightarrow{P} \mu_0 - \mu_1,$$

$$\frac{C_{0-}\widehat{\pi}_{0|0} + C_{1-}\widehat{\pi}_{0|1}}{C_{0-}} = \frac{\frac{C_{0-}}{n_1-N_2}\widehat{\pi}_{0|0} + \frac{C_{1-}}{n_1-N_2}\widehat{\pi}_{0|1}}{\frac{C_{0-}}{n_1-N_2}} \xrightarrow{P} \frac{\pi_0\pi_{0|0} + \pi_1\pi_{0|1}}{\pi_0} = \frac{p_0}{\pi_0}.$$

From this and $\mu_0 - \mu_1 = \pi_0 E\{A_{t0} - A_{t1}|A_{t0} \neq A_{t1}\)$, it follows that

$$\widehat{\mu}_{2w} \xrightarrow{P} \mu_0 - \frac{N_2}{n_1}(\mu_0 - \mu_2) - \frac{n_1 - N_2}{n_1} p_0 E\{A_{t0} - A_{t1}|A_{t0} \neq A_{t1}\).$$

Only if $E\{A_{t0} - A_{t1}|A_{t0} \neq A_{t1}\} = E\{A_{t0} - A_{t2}|A_{t0} \neq A_{t2}\}$, we have $p_0 E\{A_{t0} - A_{t1}|A_{t0} \neq A_{t1}) = (p_0 E\{A_{t0} - A_{t2}|A_{t0} \neq A_{t2}\} =)\mu_0 - \mu_2$ and hence $\widehat{\mu}_{2w} \xrightarrow{P} \mu_2$. □

6.4.4 A simulation study

We compare the performance of the estimators of this section by simulation. The simulation procedure we use is almost identical to the one of Barnett *et al.* (2001) Section 5.

The simulations (runsize 10,000) are performed for several sets of given classification probabilities and sample sizes; see Table 6.4.1. The n_1 book values are drawn from the following distribution:

book value	100	500	1000	2000	5000
probability	0.9	0.05	0.03	0.015	0.005

.

The classifications of the items are drawn from multinomial distributions. The

fractional error sizes have the following uniform distributions:

$$\begin{aligned}
&\frac{A_{t0} - A_{t1}}{A_{t0}} \sim U(0,1), && \text{if } A_{t0} \neq A_{t1},\\
&\frac{A_{t0} - A_{t2}}{A_{t0}} \sim U(0,1), && \text{if } A_{t0} = A_{t1}, A_{t0} \neq A_{t2},\\
&\frac{A_{t0} - A_{t2}}{A_{t0}} = 1 - \frac{A_{t1}}{A_{t0}}, && \text{if } A_{t0} \neq A_{t1}, A_{t0} \neq A_{t2}, A_{t1} \neq A_{t2}.
\end{aligned}$$

So far the simulation procedure is identical to the one of Barnett *et al.* (2001). However, to avoid not uniquely defined parameter estimates (see Section 3.3.3), we apply stratified sampling instead of random sampling (see Section 3.2.3).

From the described simulation procedure, the mean population error size can be determined analytically for each set of classification probabilities. In each simulation run $\mu_0 - \mu_2$ is estimated using the four discussed estimators. Note that $E\{A_{t0} - A_{t1} | A_{t0} \neq A_{t1}\} = E\{A_{t0} - A_{t2} | A_{t0} \neq A_{t2}\}$ in the described simulation procedure. Table 6.4.1 contains the results of the simulations.

From the four studied estimators, $\widehat{\mu}_{2r}$ has the largest bias; the other three estimators have a small bias (if any at all). The small bias of $\widehat{\mu}_{2w}$ (never exceeding 0.1) is caused by the fact that $E\{A_{t0} - A_{t1} | A_{t0} \neq A_{t1}\} = E\{A_{t0} - A_{t2} | A_{t0} \neq A_{t2}\}$ for the simulated data.

Higher sample sizes in the first and second round lead to a lower variance for all estimators except $\widehat{\mu}_{2p}$; the variance of $\widehat{\mu}_{2p}$ decreases for higher n_2, but n_1 hardly seems to have an impact. See for example the first entry of the second half of the table: the standard deviation of $\widehat{\mu}_{2p}$ is 11.9, 12.0 and 7.0 for (n_1, n_2) equal to (1000,100), (3000,100) and (3000,300), respectively.

We see that the variances of all estimators are lower for the small mean error size (10) than for the high mean error size (20). For example, for $n_1 = 1000$ and $n_2 = 100$ the standard deviation of $\widehat{\mu}_2$ is 3.1 for the first set of probability parameters with $\mu_0 - \mu_2 = 10$; for the first set of parameter values with $\mu_0 - \mu_2 = 20$ the standard deviation is 4.1.

In every second line of the table the probability of an auditor missing an error is higher, and the probability of an auditor finding the right size of an error is lower than in the previous line. Comparing two subsequent lines, we see that a

Probabilities					$n_1 = 1000$ and $n_2 = 100$				$n_1 = 3000$ and $n_2 = 100$				$n_1 = 3000$ and $n_2 = 300$			
π_{11}	π_{10}	π_{01}	π_{00e}	π_{00u}	$\widehat{\mu}_2$	$\widehat{\mu}_{2r}$	$\widehat{\mu}_{2p}$	$\widehat{\mu}_{2w}$	$\widehat{\mu}_2$	$\widehat{\mu}_{2r}$	$\widehat{\mu}_{2p}$	$\widehat{\mu}_{2w}$	$\widehat{\mu}_2$	$\widehat{\mu}_{2r}$	$\widehat{\mu}_{2p}$	$\widehat{\mu}_{2w}$
Mean error size = 10																
.89	.02	.01	.06	.02	10.1	10.1	9.9	10.0	10.0	10.1	9.9	10.0	10.0	10.1	10.1	10.0
					(3.1)	(6.3)	(8.7)	(3.4)	(2.4)	(6.0)	(8.5)	(2.4)	(1.7)	(3.8)	(5.0)	(1.9)
.89	.06	.01	.02	.02	10.0	10.0	10.0	9.9	10.0	10.2	10.1	10.0	10.0	10.2	10.1	10.0
					(3.8)	(7.9)	(9.1)	(4.5)	(3.5)	(8.2)	(9.1)	(3.4)	(2.3)	(4.9)	(5.0)	(2.6)
.87	.02	.03	.06	.02	10.0	10.2	9.9	10.0	10.0	10.3	9.9	10.0	10.0	10.1	9.9	10.0
					(3.1)	(6.8)	(8.7)	(3.4)	(2.7)	(6.8)	(8.8)	(2.6)	(1.8)	(4.2)	(5.0)	(1.9)
.87	.06	.03	.02	.02	10.1	10.3	10.0	10.0	10.1	10.3	10.0	10.0	10.0	10.1	9.9	10.0
					(3.8)	(8.9)	(8.8)	(4.2)	(3.5)	(8.6)	(8.7)	(3.3)	(2.4)	(5.4)	(5.0)	(2.4)
.85	.02	.05	.06	.02	10.1	10.4	10.1	10.0	10.0	10.3	10.0	10.0	10.0	10.2	10.0	10.0
					(3.3)	(7.7)	(9.0)	(3.4)	(2.8)	(7.6)	(8.8)	(2.7)	(1.8)	(4.7)	(5.1)	(1.9)
.85	.06	.05	.02	.02	10.0	10.3	10.0	10.0	10.0	10.4	10.1	10.0	10.0	10.4	10.1	10.0
					(3.8)	(9.4)	(8.8)	(4.0)	(3.6)	(9.5)	(9.0)	(3.3)	(2.4)	(5.8)	(5.1)	(2.3)
Mean error size = 20																
.78	.04	.02	.12	.04	20.0	20.2	19.8	20.0	20.1	20.2	19.9	20.1	20.0	20.1	19.9	20.0
					(4.1)	(8.8)	(11.9)	(4.7)	(3.3)	(8.5)	(12.0)	(3.3)	(2.5)	(5.3)	(7.0)	(2.7)
.78	.12	.02	.04	.04	20.0	20.3	20.1	20.0	20.0	20.3	20.0	20.0	20.1	20.2	20.1	20.0
					(5.4)	(11.5)	(12.2)	(6.3)	(5.1)	(11.4)	(12.4)	(4.7)	(3.5)	(7.0)	(7.2)	(3.6)
.74	.04	.06	.12	.04	19.9	20.2	19.9	20.0	20.0	20.4	20.1	20.0	20.0	20.3	20.0	20.0
					(4.2)	(9.9)	(12.2)	(4.6)	(3.6)	(10.1)	(12.5)	(3.6)	(2.6)	(6.1)	(6.9)	(2.7)
.74	.12	.06	.04	.04	20.0	20.4	20.0	20.0	20.0	20.5	19.9	20.0	20.0	20.3	20.0	20.0
					(5.5)	(12.5)	(12.3)	(5.9)	(5.2)	(12.3)	(12.5)	(4.6)	(3.6)	(7.6)	(7.2)	(3.3)
.70	.04	.06	.12	.04	20.0	20.6	20.1	20.0	20.0	20.4	19.9	20.0	20.0	20.4	20.0	20.0
					(4.4)	(10.9)	(12.6)	(4.7)	(3.8)	(10.7)	(12.2)	(3.7)	(2.6)	(6.8)	(7.1)	(2.7)
.74	.12	.06	.04	.04	20.0	20.5	19.9	20.0	20.1	20.9	20.2	20.0	20.0	20.4	20.0	20.0
					(5.5)	(13.7)	(12.3)	(5.6)	(5.3)	(13.5)	(12.7)	(4.6)	(3.5)	(8.1)	(7.2)	(3.2)

Table 6.4.1: Simulated means (and standard deviations) of the estimators

higher π_{10} and a lower π_{00e} cause an increase in the variance of the estimators. For example, in the first two lines of the table the standard deviation of $\widehat{\mu}_2$ increases from 3.1 to 3.8 for $n_1 = 1000$ and $n_2 = 100$.

Based on the results of Table 6.4.1, we can conclude that estimators $\widehat{\mu}_2$ and $\widehat{\mu}_{2w}$ have comparable variances and outperform $\widehat{\mu}_{2r}$ and $\widehat{\mu}_{2p}$ (in terms of variance). The simulations in this section were constructed such that $E\{A_{t0} - A_{t1}|A_{t0} \neq A_{t1}\} = E\{A_{t0} - A_{t2}|A_{t0} \neq A_{t2}\}$, which is a necessary condition for consistency of $\widehat{\mu}_{2w}$. This is not an essential condition for the consistency of $\widehat{\mu}_2$. Moreover, $\widehat{\mu}_{2w}$ does not outperform $\widehat{\mu}_2$ even under this condition and with a model for the simulated data which deviates from our model in Section 6.2. Hence, $\widehat{\mu}_2$ seems to be the preferable estimator.

6.5 Final remarks and conclusions

We introduced a mixed model for a repeated audit control with two rounds. This model consists of a submodel for the absolute classification probabilities and another submodel in terms of conditional regression for the audit values. The generalization to a repeated audit control with k rounds is quite straightforward. The basic variables of the general model are $A_0, A_1, \ldots, A_k$, where A_i $(i = 1, \ldots, k)$ is the value according to auditor i of a random record. The records can be classified based on the question whether some of the k audit values and book values coincide; note that the number of classifications increases sharply in k. Next, similar to Section 6.2.3, conditional regression models can be specified for the audit values which do not coincide with the book value or previous audit values according to the classification.

As mentioned previously, repeated audit controls can be regarded as a missing data problem (or more specific: as a monotone missing data problem). In the missing data literature, Olkin and Tate (1961) have already introduced a model with a mixture of both categorical and continuous variables: the general location model. In this model, K categorical variables are classified, and the M continuous variables have a (M-variate) normal distribution conditional on this classification. The model in this chapter differs essentially from the general location model: the

classifications are not based on separate categorical variables but on the equality of the continuous variables, and the dimensionality of the conditional models may be lower than M. For example, the conditional regression models in Table 6.2.2 are uni- and bivariate.

We derived estimators for the model parameters and the main parameter of interest: the mean true value. In a simulation study our estimator for the mean true value outperformed several other estimators introduced by Barnett *et al*. (2001), although the underlying model of the simulation study differed from our model in Section 6.2.

So far we have only discussed point estimators for the parameters, but confidence limits are at least as important in auditing practice. In auditing practice, selection with probabilities proportional to the recorded value ('monetary unit sampling') is applied frequently instead of the discussed sampling techniques. It would be interesting to investigate this sampling method as well. We leave these topics for further research.

6.6 Appendices

6.6.1 Estimators for the regression parameters

We use the following notation for sample averages and variances:

$$\begin{aligned}\overline{A}_g^{(C_{ij})} &= \frac{1}{C_{ij}}\sum^{C_{ij}} A_{tg},\\ \overline{S}_{gh}^{(C_{ij})} &= \frac{1}{C_{ij}}\sum^{C_{ij}}(A_{tg}-\overline{A}_g^{(C_{ij})})(A_{th}-\overline{A}_h^{(C_{ij})}).\end{aligned}$$

OLS-estimators

$$b_1 = \begin{bmatrix} \overline{A}_2^{(C_{10})} - (S_{00}^{(C_{10})})^{-1}S_{02}^{(C_{10})}\overline{A}_0^{(C_{10})} \\ (S_{00}^{(C_{10})})^{-1}S_{02}^{(C_{10})} \end{bmatrix}, \quad s_1^2 = \frac{\sum^{C_{10}}(A_{t2} - b_1'\begin{bmatrix} 1 \\ A_{t0}\end{bmatrix})^2}{C_{10}-2}$$

$$b_0 = \begin{bmatrix} \overline{A}_1^{(C_0)} - (S_{00}^{(C_0)})^{-1}S_{01}^{(C_0)}\overline{A}_0^{(C_0)} \\ (S_{00}^{(C_0)})^{-1}S_{01}^{(C_0)} \end{bmatrix}, \qquad s_0^2 = \frac{\sum^{C_0}(A_{t1} - b_0'\begin{bmatrix} 1 \\ A_{t0}\end{bmatrix})^2}{C_0-2}$$

$$b_{0u} = \begin{bmatrix} \overline{A}_2^{(C_{00u})} - (S_{00}^{(C_{00u})})^{-1}S_{02}^{(C_{00u})}\overline{A}_0^{(C_{00u})} \\ (S_{00}^{(C_{00u})})^{-1}S_{02}^{(C_{00u})} \end{bmatrix}$$

$$s_{0u}^2 = \frac{\sum^{C_{00u}}(A_{t2} - b_{0u}'\begin{bmatrix} 1 \\ A_{t0}\end{bmatrix})^2}{C_{00u}-2}$$

$$s_{12} = \frac{1}{C_{00u}-2}\sum^{C_{00u}}(A_{t1} - b_0'\begin{bmatrix} 1 \\ A_{t0}\end{bmatrix})(A_{t2} - b_{0u}'\begin{bmatrix} 1 \\ A_{t0}\end{bmatrix})$$

ML estimators

$$\widehat{\beta}_1 = b_1, \qquad \widehat{\sigma}_1^2 = \frac{C_{10}-2}{C_{10}} s_1^2, \qquad \widehat{\beta}_0 = b_0, \qquad \widehat{\sigma}_0^2 = \frac{C_0-2}{C_0} s_0^2$$

$$\begin{bmatrix} \widehat{\beta}_{0u} \\ \widehat{\alpha}_{0u} \end{bmatrix} = \begin{bmatrix} C_{00u} & \sum^{C_{00u}} A_{t0} & \sum^{C_{00u}} \widehat{\varepsilon}_{t1} \\ \sum^{C_{00u}} A_{t0} & \sum^{C_{00u}} A_{t0}^2 & \sum^{C_{00u}} A_{t0}\widehat{\varepsilon}_{t1} \\ \sum^{C_{00u}} \widehat{\varepsilon}_{t1} & \sum^{C_{00u}} A_{t0}\widehat{\varepsilon}_{t1} & \sum^{C_{00u}} \widehat{\varepsilon}_{t1}^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum^{C_{00u}} A_{t2} \\ \sum^{C_{00u}} A_{t0}A_{t2} \\ \sum^{C_{00u}} \widehat{\varepsilon}_{t1}A_{t2} \end{bmatrix}$$

$$\widehat{\sigma}_{0u}^2 = \widehat{\sigma}_0^2\widehat{\alpha}_{0u}^2 + \frac{1}{C_{00u}} \sum^{C_{00u}} (A_{t2} - \widehat{\beta}_{0u}' \begin{bmatrix} 1 \\ A_{t0} \end{bmatrix} - \alpha_{0u}\widehat{\varepsilon}_{t1})^2, \quad \widehat{\sigma}_{12} = \widehat{\sigma}_0^2\widehat{\alpha}_{0u},$$

where $\widehat{\varepsilon}_{t1} = A_{t1} - \widehat{\beta}_0' \begin{bmatrix} 1 \\ A_{t0} \end{bmatrix}$.

6.6.2 Conditional expectations

$$E\{A_{t2}|A_{t0}, A_{t1}, A_{t0} = A_{t1}, \widehat{\theta}\} = \widehat{\pi}_{1|1}A_{t0} + \widehat{\pi}_{0|1}\widehat{\beta}_0' \begin{bmatrix} 1 \\ A_{t0} \end{bmatrix}$$

$$E\{A_{t2}|A_{t0}, A_{t1}, A_{t0} \neq A_{t1}, \widehat{\theta}\} = \widehat{\pi}_{1|0}A_{t0} + \widehat{\pi}_{0e|0}A_{t1} + \widehat{\pi}_{0u|0}(\widehat{\beta}_{0u}' \begin{bmatrix} 1 \\ A_{t0} \end{bmatrix} + \widehat{\alpha}_{0u}\widehat{\varepsilon}_{t1})$$

$$E\{A_{t2}|A_{t0}, \widehat{\theta}\} = (\widehat{\pi}_{11} + \widehat{\pi}_{01})A_{t0} + \widehat{\pi}_{10}\widehat{\beta}_1' \begin{bmatrix} 1 \\ A_{t0} \end{bmatrix} + \widehat{\pi}_{00e}\widehat{\beta}_0' \begin{bmatrix} 1 \\ A_{t0} \end{bmatrix} + \widehat{\pi}_{00u}\widehat{\beta}_{0u}' \begin{bmatrix} 1 \\ A_{t0} \end{bmatrix}$$

Bibliography

Afifi, A.A. and R.M. Elashoff (1966). Missing Observations in Multivariate Statistics: I. Review of the Literature. *Journal of the American Statistical Association*, **61**, 595–604.

Anderson, T.W. (1957). Maximum Likelihood Estimates for a Multivariate Normal Distribution when Some Observations are Misssing. *Journal of the American Statistical Association*, **52**, 200–203.

Barnett, V., J. Haworth, and T.M.F. Smith (2000). A Two-Phase Sampling Scheme with Applications to Auditing. Technical Report 2, Department of Mathematics, Statistics and Operational Research, Nottingham Trent University.

Barnett, V., J. Haworth, and T.M.F. Smith (2001). A Two-Phase Sampling Scheme with Applications to Auditing or Sed Quis Custodiet Ipsos Custodes. *Journal of the Royal Statistical Society A*, **164**, 407–422.

Bartlett, M.S. (1947). Multivariate Analysis. *Journal of the Royal Statistical Society B*, **9**, 176–197.

Bhargava, R.P. (1962). *Multivariate Tests of Hypotheses with Incomplete Data*. Ph.d. dissertation, Stanford University.

Bishop, Y.M.M., S.E. Fienberg, and P.W. Holland (1975). *Discrete Multivariate Analysis: Theory and Practice*. MA:MIT press.

Box, G.E.P. (1949). A General Distribution Theory for a Class of Likelihood Criteria. *Biometrika*, **36**, 317–346.

Cox, D.R. and D.V. Hinkley (1974). *Theoretical Statistics*. Chapman and Hall.

Cox, D.R. and E.J. Snell (1979). On Sampling and the Estimation of Rare Errors. *Biometrika*, **66**, 125–132.

Dempster, A.P., N.M. Laird, and D.B. Rubin (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society B*, **39**, 1–38.

Fienberg, S.E., J. Neter, and R.A. Leitch (1977). Estimating the Total Overstatement Error in Accounting Populations. *Journal of the American Statistical Association*, **72**, 295–302.

Fujisawa, H. (1995). A Note on the Maximum Likelihood Estimators for Multivariate Normal Distribution with Monotone Data. *Communications in Statistics: Theory and Methods*, **24**, 1377–1382.

Geng, Z. and C. Asano (1989). Bayesian Estimation Methods for Categorical Data with Misclassifications. *Communications in Statistics: Theory and Methods*, **18**, 2935–2954.

Genugten, B.B. van der (1988). *Inleiding tot de Waarschijnlijkheidsrekening en Mathematische Statistiek, deel 2*. Stenfert Kroese (In Dutch).

Genugten, B.B. van der (1997). Testing in the Restricted Linear Model Using Canonical Partitions. *Linear Algebra and its Applications*, **264**, 349–353.

Green, W.E. (1993). *Econometric analysis* (second ed.). Prentice-Hall.

Gunel, E. (1984). A Bayesian Analysis of the Multinomial Model for a Dichotomous Response with Nonrespondents. *Communications in Statistics: Theory and Methods*, **13**, 737–751.

Gupta, A.K. and D.K. Nagar (2000). *Matrix Variate Distributions*. Chapman and Hall/CRC.

Hao, J. and K. Krishnamoorthy (2001). Inferences on a Normal Covariance Matrix and Generalized Variance with Monotone Missing Data. *Journal of Multivariate Analysis*, **78**, 62–82.

Jinadasa, K.G. and D.S. Tracy (1992). Maximum Likelihood Estimation for Multivariate Normal Distribution with Monotone sample. *Communications*

in Statistics: Theory and Methods, **21**, 41–50.

Johnson, J.R., R.A. Leitch, and J. Neter (1981). Characteristics of Errors in Accounts Receivable and Inventory Audits. *The Accounting Review*, **56**, 270–293.

Kanda, T. and Y. Fujikoshi (1998). Some Basic Properties of the MLE's for a Multivariate Normal Distribution with Monotone Missing Data. *American Journal of Mathematical and Management Sciences*, **18**, 161–190.

Kres, H. (1983). *Statistical Tables for Multivariate Analysis*. Springer-Verlag.

Krishnamoorthy, K. (1991). Estimation of Normal Covariance and Precision Matrices with Incomplete Data. *Communications in Statistics: Theory and Methods*, **20**, 757–770.

Krishnamoorthy, K. and M.K. Pannala (1998). Some Simple Tests Procedures for Normal Mean Vector with Incomplete Data. *Annals of the Institute of Statistical Mathematics*, **50**, 531–542.

Krishnamoorthy, K. and M.K. Pannala (1999). Confidence Estimation of a Normal Mean Vector with Incomplete Data. *Canadian Journal of Statistics*, **27**, 395–407.

Laws, D.J. and A. O'Hagan (2000). Bayesian Inference for Rare Errors in Populations with Unequal Unit Sizes. *Applied Statistics*, **49**, 577–590.

Lehmann, E.L. and G. Casella (1998). *Theory of Point Estimation*. John Wiley.

Little, R.J.A. and D.B. Rubin (2002). *Statistical Analysis with Missing Data*. John Wiley and Sons.

Magnus, J.R. (1978). Maximum Likelihood Estimation of the GLS Model with Unknown Parameters in the Disturbance Covariance Matrix. *Journal of Econometrics*, **7**, 281–312.

Malinvaud, E. (1970). *Statistical Methods of Econometrics*. Elsevier.

McLachan, G.J. and T. Krishnan (1997). *The EM Algorithm and Extensions*. John Wiley and Sons.

Meng, X.L. and D.B. Rubin (1993). Maximum Likelihood Estimation via the ECM Algorithm: A General Framework. *Biometrika*, **80**, 267–278.

Mittelhammer, R.C., G.G. Judge, and D.J. Miller (1996). *Econometric Foundations*. University Press.

Moors, J.J.A. (1983). Bayes' Estimation in Sampling for Auditing. *Statistician*, **32**, 281–288.

Moors, J.J.A. (1999). Double Checking for Two Error Types. CentER Discussion Paper 23, Tilburg University.

Moors, J.J.A., B.B. van der Genugten, and L.W.G. Strijbosch (2000). Repeated Audit Controls. *Statistica Neerlandica*, **54**, 3–13.

Moors, J.J.A. and M.J.B.T Janssens (1989). Exact Distributions of Bayesian Cox-Snell Bounds in Auditing. *Journal of Accounting Research*, **27**, 135–144.

Muirhead, R.J. (1982). *Aspects of Multivariate Statistical Theory*. Wiley.

Neter, J., J.R. Johnson, and R.A. Leitch (1985). Characteristic of Dollar-Unit Taints and Error Rates in Accounts Receivable and Inventory. *The Accounting Review*, **60**, 488–499.

Olkin, I. and R.F. Tate (1961). Multivariate Correlation Models with Mixed Discrete and Continuous Variables. *Annals of Mathematical Statistics*, **32**, 448–465.

Raats, V.M. (1999). *Herhaalde Steekproefcontrole*. Master's thesis, Tilburg University (In Dutch).

Raats, V.M. (2004). Approximations of the Generalized Wilks' Distribution. CentER Discussion Paper 85, Tilburg University.

Raats, V.M., B.B. van der Genugten, and J.J.A. Moors (2002a). A General Model for Repeated Audit Controls Using Monotone Subsampling. CentER Discussion Paper 10, Tilburg University.

Raats, V.M., B.B. van der Genugten, and J.J.A. Moors (2002b). Multivariate Regression with Monotone Missing Observations of the Dependent Variables. CentER Discussion Paper 63, Tilburg University.

Raats, V.M., B.B. van der Genugten, and J.J.A. Moors (2004a). Asymptotics of Multivariate Regression with Consequetively Added Dependent Variables. CentER Discussion Paper 77, Tilburg University.

Raats, V.M., B.B. van der Genugten, and J.J.A. Moors (2004b). A General Model for Repeated Audit Controls Using Monotone Subsampling. *Communications in Statistics: Theory and Methods*, **33**, 949–977.

Raats, V.M. and J.J.A. Moors (2000). Double checking for two error types. CentER Discussion Paper 120, Tilburg University.

Raats, V.M. and J.J.A. Moors (2003). Double Checking Auditors: a Bayesian Approach. *Statistician*, **52**, 1–15.

Raats, V.M., J.J.A. Moors, and B.B. van der Genugten (2004). A Mixed Model for Double Checking Fallible Auditors. CentER Discussion Paper 82, Tilburg University.

Rao, C.R. (1952). *Advanced Statistical Methods in Biometric Research*. Wiley.

Rencher, A.C. (1998). *Multivariate Statistical Inference and Applications*. John Wiley and Sons.

Robins, J.M. and A. Rotnitzky (1995). Semiparametric Efficiency in Multivariate Regression Models with Missing Data. *Journal of the American Statistical Association*, **90**, 122–129.

Rubin, D.B. (1976). Inference and Missing Data. *Biometrika*, **63**, 581–592.

Schafer, J.L. (1997). *Analysis of Incomplete Multivariate Data*. Chapmann and Hall.

Seber, G.A.F. (1984). *Multivariate Observations*. John Wiley.

Srivastava, M.S. (1985). Multivariate Data with Missing Observations. *Communications in Statistics: Theory and Methods*, **14**, 775–792.

Tamura, H. (1988). Estimation of Rare Errors Using Expert Judgement. *Biometrika*, **75**, 1–9.

Tamura, H. and P.A. Frost (1986). Tightening CAV (DUS) Bounds by Using a Parametric Model. *Journal of Accounting Research*, **24**, 364–371.

Tanner, M.A. and W.H. Wong (1987). The Calculation of Posterior Distributions by Data Augmentation. *Journal of the American Statistical Association*, **82**, 528–550.

Tenenbein, A. (1970). A Double Sampling Scheme for Estimating from Binomial Data with Misclassifications. *Journal of the American Statistical Association*, **65**, 1350–1361.

Tenenbein, A. (1971). A Double Sampling Scheme for Estimating from Binomial Data with Misclassifications; Sample Size Determination. *Biometrics*, **27**, 935–944.

Tenenbein, A. (1972). A Double Sampling Scheme for Estimating from Inspection. *Technometrics*, **14**, 187–202.

Viana, M.A.G. (1994). Bayesian Small-Sample Estimation of Misclassified Multinomial Data. *Biometrics*, **50**, 237–243.

Wu, C.J.F. (1983). On the Convergence Properties of the EM Algorithm. *Annals of Statistics*, **11**, 95–103.

York, J., D. Madigan, I. Heuch, and R.T. Lie (1995). Birth Defects Registered by Double Sampling: a Bayesian Approach Incorporating Covariates and Model Uncertainty. *Applied Statistics*, **44**, 227–242.

Samenvatting (Summary in Dutch)

Statistische modellen voor steekproefcontroles zijn meestal gebaseerd op de (impliciete) veronderstelling dat de controleur geen fouten maakt. Echter, controleurs zijn menselijk en dus feilbaar.

Eén manier om rekening te houden met mogelijke fouten van een controleur is het toepassen van een herhaalde steekproefcontrole. Een herhaalde steekproefcontrole bestaat uit twee of meer ronden. In de eerste ronde worden posten uit de boekhouding steekproefsgewijs gecontroleerd door een feilbare controleur. In de daaropvolgende ronde wordt een deelsteekproef van deze posten nogmaals gecontroleerd, ditmaal door een meer bekwame controleur. Dit kan enkele malen herhaald worden totdat de laatste controleur, een feilloze expert, de juiste waarde geeft voor een deelsteekproef van posten die door alle voorgaande (feilbare) controleurs al gecontroleerd zijn.

Herhaalde steekproefcontroles zijn gerelateerd aan ontbrekende data problemen. Standaard statistische methoden analyseren meestal data van een aantal variabelen, waargenomen voor een vast aantal cases. Het komt vaak voor dat voor enkele cases niet alle variabelen zijn waargenomen, zodat enkele observaties ontbreken. Deze ontbrekende dataproblemen zijn uitgebreid in de literatuur bestudeerd. Herhaalde steekproefcontroles kunnen beschouwd worden als ontbrekende data problemen. Neem bijvoorbeeld de herhaalde steekproefcontrole met twee ronden: het oordeel van de expert is slechts beschikbaar voor de dubbel gecontroleerde steekproefposten, maar ontbreekt voor de eenmalig gecontroleerde posten.

Dit proefschrift behandelt de statistische modellering en analyse van herhaalde steekproefcontroles. De modellen verschillen met betrekking tot het aantal feil-

bare controleurs en het soort variabelen (categorisch, continu of een combinatie van beide). Hoofdstuk 2 behandelt de modellering en analyse van de meest eenvoudige situatie met één feilbare controleur en alternatieve variabelen; dat laatste wil zeggen dat de controleur en expert de posten slechts als correct dan wel incorrect classificeren. Het model van Hoofdstuk 2 is al eerder beschreven in de literatuur, maar de aandacht is tot nu toe voornamelijk uitgegaan naar puntschattingen voor de fractie incorrecte posten in de hele boekhouding. Aangezien bovengrenzen in de praktijk vaak minstens zo belangrijk zijn als puntschattingen, bespreken we twee methoden voor het bepalen van bovengrenzen: de zogenaamde klassieke methode en de Bayesiaanse methode. Het verschil is dat de Bayesiaanse methode gebruik maakt van eventueel aanwezige (subjectieve) voorkennis omtrent de populatie en de kwaliteit van de controleurs. De klassieke methode blijkt te leiden tot erg hoge betrouwbaarheidsbovengrenzen; de Bayesiaanse aanpak geeft in het algemeen lagere bovengrenzen.

In Hoofdstuk 3 presenteren we een algemeen kader voor herhaalde steekproeven; er kan meer dan één feilbare controleur bij betrokken zijn en bovendien beschouwen we categorische variabelen: er kunnen meer classificatiemogelijkheden zijn dan alleen correct en incorrect. Het model van het voorgaande hoofdstuk is hiervan dus het meest eenvoudige geval. We bespreken twee verschillende methoden voor het trekken van de steekproefposten. Voor beide steekproefmethoden bepalen we de meest aannemelijke schatters en geven we een oplossing voor het probleem van niet uniek bepaalde schatters. We vergelijken ook drie verschillende methoden voor het bepalen van bovengrenzen, waaronder de Bayesiaanse aanpak. Ons Bayesiaans model verschilt van het gangbare in de wijze waarop we de voorkennis formuleren.

In de laatste drie hoofdstukken bespreken we modellen voor continue variabelen of een combinatie van categorische en continue. Hoofstukken 4 en 5 behandelen multivariate lineaire regressie met een monotone datastructuur voor de afhankelijke variabelen. In multivariate regressie wordt een aantal afhankelijke variabelen beschreven met behulp van een aantal verklarende variabelen. Een monotone datastructuur voor de afhankelijke (continue) variabelen betekent het volgende: de verklarende variabelen kunnen zodanig geordend worden dat als een waarne-

ming van een verklarende variabele ontbreekt voor een case, dan ontbreken ook de waarnemingen van alle daaropvolgende verklarende variabelen voor dezelfde case. De waarnemingen voor de verklarende variabelen zijn compleet. Een zeer speciaal geval is het model met slechts de constante term als verklarende variabele dat al uitvoerig in de literatuur besproken is.

In Hoofdstuk 4 bepalen we analytische uitdrukkingen voor enkele schatters door middel van projecties; deze schatters hebben een duidelijke meetkundige interpretatie. Voor het bepalen van schattingen wordt in ontbrekende data problemen vaak gebruik gemaakt van een iteratief algoritme; dit zogenaamde EM-algoritme convergeert numeriek naar de meest aannemelijke schattingen. In vergelijking hiermee, heeft onze methode twee voordelen: de gemakkelijke interpretatie en de directe berekening die natuurlijk nauwkeuriger en sneller is. We bespreken ook in detail een toets voor de regressiecoefficienten: de zogenaamde *likelihood ratio test*. De toetsingsgrootheid wordt afgeleid, alsmede de bijbehorende kansverdeling, die een generalisatie van reeds bestaande kansverdelingen is. Voor deze nieuwe kansverdeling worden verschillende benaderingen afgeleid en vergeleken door middel van simulatie.

In Hoofdstuk 5 komen verschillende aspecten van het multivariate regressiemodel aan de orde. We laten zien dat de schatters van het vorige hoofdstuk consistent zijn, dit wil zeggen dat het verschil tussen de schatters en de parameters naar nul gaat voor grote steekproeven. Voor de volledigheid worden ook twee alternatieve schattingsmethoden gegeven voor het bepalen van de meest aannemelijke schatters; beide methoden zijn veelgebruikte iteratieve algoritmes die numeriek convergeren naar de meest aannemelijke schattingen. Ten slotte bekijken we ook een generalisatie van het model met slechts de constante als verklarende variabele: *one-way* MANOVA.

In de praktijk is men vaak geïnteresseerd in de totale grootte van fouten in de populatie; in geval van bekende populatie-omvang is dit equivalent aan de gemiddelde grootte van de fouten. De fout bij de meeste posten is echter gelijk aan nul, zodat het niet realistisch is een continu model voor de grootte van de fouten te veronderstellen. In Hoofdstuk 6 construeren we een realistischer model voor de grootte van fouten door de modellen van de voorgaande hoofdstukken te com-

bineren. Voor de classificatiekansen gebruiken we de modellen van Hoofdstukken 2 en 3. Als uit de classificatie van een post vervolgens blijkt dat er echt sprake is van een fout, dan wordt de grootte van deze fout gemodelleerd met behulp van een conditioneel regressiemodel (vergelijkbaar met dat van Hoofdstuk 4). De schatters voor de modelparameters en voor de gemiddelde grootte van de fouten in de populatie zijn nu eenvoudig te bepalen door combinatie van de schattingstechnieken van de voorgaande hoofdstukkken. Simulatie toont aan dat onze schatter voor de gemiddelde grootte van de fouten nauwkeuriger is dan enkele andere schatters die eerder in de literatuur besproken zijn.